TECHNISCHE
UNIVERSITÄT
DARMSTADT

Three Essays in Energy and Environmental Economics

Dissertation

Submitted in fulfillment of the requirements for the degree of

doctor rerum politicarum (Dr. rer. pol.)

by the Department of Law and Economics of the
Technical University of Darmstadt

by

Moritz Tarach, M.Sc.
(Place of birth: Hamburg)

First supervisor: Prof. Dr. Jens Krüger
Second supervisor: Prof. Dr. Michael Neugart
Filing date: 29.05.2024
Date of defense: 05.11.2024

Place of publication: Darmstadt
Year of the viva voce: 2024

D 17

Tarach, Moritz
Three Essays in Energy and Environmental Economics
Darmstadt, Technische Universität Darmstadt
Year thesis published in TUprints: 2025
Date of defense: 05.11.2024

Acknowledgements

This dissertation was written while I was research and teaching assistant at the Chair of Empirical Economics at the Technical University of Darmstadt, Germany. Without the support of many people, whom I sincerely thank in the following, the completion of this dissertation would not have been possible.

First and foremost, I am deeply indebted to my supervisor, Prof. Dr. Jens Krüger for giving me the opportunity to pursue my doctoral thesis. Working with him has been a pleasure which I greatly appreciate. I want to thank him for his invaluable advice, constructive support, and guidance while writing this dissertation. I am thankful to him for sharing his academic experience with me, especially his knowledge in writing and publishing academic papers.

Moreover, I would like to thank Prof. Dr. Michael Neugart for his review as my second supervisor. I would also like to thank Prof. Dr. Dirk Schiereck for chairing the examination committee as well as the committee member Prof. Frank Pisch for their questions and insights shared during my defense.

Furthermore, I thank my (former) colleagues from the Technical University of Darmstadt for the pleasant and cooperative working atmosphere. I am especially grateful to my colleagues from the Chair of Empirical Economics: Vanessa Belew, Dr. Jingwei Pan, Dr. Seulki Chung, Yannick Schmidt and Fahim Safi. I am also grateful to my colleagues from neighboring chairs, particularly from the chairs 'Technology and Innovation Management' and 'Public Economics and Economic Policy', for spending time together in a friendly and cheerful atmosphere during lunch breaks.

Besides, I am thankful to project partners at the University of Kassel, Prof. Dr. Heike Wetzel and Larissa Fait, for their efforts in compiling the data that I use in Chapter 3 of this dissertation.

Last but not least, I want to thank my friends and my family, especially my parents and my brother. Without their unwavering support and inspiration, during the years of writing this dissertation and before during my studies, the completion of this dissertation would not have been possible.

Abstract

This dissertation consists of three studies, each examining a different topic in the field of energy and environmental economics. The topics comprise: (i) estimating potentials for greenhouse gas emission reductions of economic sectors, (ii) forecasting the oil production of a region based on historical data from discoveries, and (iii) examining the determinants of electricity price fluctuations. Each of the studies uses a particular statistical method or mathematical model that is specifically adapted to the research question and the data set under investigation.

The first study is a stochastic nonparametric efficiency analysis in which greenhouse gas emissions are included as bad outputs. For seven economic sectors and sixteen European countries, this study estimates greenhouse gas emission reduction potentials, i.e., the quantity of emissions that could potentially be reduced by improvements in productive efficiency. The standard DEA method is extended by a specific bootstrapping procedure used to implement a bias correction and to compute confidence intervals. The magnitudes of the emission reduction potentials are compared with the emission reduction targets for 2030 from the European Commission. The results show that improvements in productive efficiency are a quantitatively important element, potentially allowing for a substantial reduction of greenhouse gas emissions in the European Union.

The second study presents a stochastic model for forecasting for an oil-producing region the amount of undiscovered oil, the future path of oil discovery and that of oil production. The model combines three submodels: (i) an empirically-founded production model at the level of individual oil fields, (ii) a successive sampling discovery model after Kaufman et al. (1975) for forecasting field sizes, and (iii) a stochastic birth process model for forecasting discovery dates. The model is estimated and evaluated for the oil-producing regions of Norway and the U.S. Gulf of Mexico (the latter further split into shallow- and deep-water parts). The results show that the predictions for oil discovery are somewhat too low compared to the actuals for Norway and for the shallow-water Gulf of Mexico, while for the deep-water Gulf of Mexico the predictions are too high. This is similarly reflected in the predictions for oil production.

The third study is a multivariate wavelet analysis of the German wholesale electricity market, which examines the determinants of electricity price fluctuations using daily time series. The possible determinants are coal prices, gas prices, and the residual load (i.e., electricity consumption minus wind and solar generation). The multivariate wavelet method allows for a detailed examination of the relations between the time series in time-frequency space, while also taking into account the interdependencies among the different time series. The results show that the residual load is the key short-run determinant of electricity prices, while coal and gas prices are the key long-run determinants. Also, this study finds that the co-movement relation among the energy prices is time-varying, which is consistent with the findings of other studies (e.g., Sousa et al. (2014); Aguiar-Conraria et al. (2018)).

Zusammenfassung

Die vorliegende Dissertation beinhaltet drei Studien, die jeweils unterschiedliche Themen aus dem Forschungsfeld der Energie- und Umweltökonomie untersuchen. Die Forschungsthemen der Studien umfassen: (i) die Schätzung von Einsparpotentialen für Treibhausgasemissionen auf Sektorebene, (ii) die Prognose der Ölförderung einer Region auf Basis von historischen Daten zu Ölfunden, (iii) die Untersuchung der Bestimmungsfaktoren von Strompreisschwankungen. Jede der Studien nutzt dabei bestimmte statistische oder mathematische Methoden, welche speziell auf die Forschungsfrage und den zu untersuchenden Datensatz zugeschnitten sind.

Die erste Studie ist eine stochastische, nichtparametrische Effizienzanalyse mit Treibhausgasemissionen als unerwünschte Outputs. Es werden für sieben Wirtschaftssektoren und sechstzehn Europäische Länder Einsparpotentiale für Treibhausgasemissionen geschätzt, d.h. die Emissionsmengen, die möglicherweise durch Produktivitätsverbesserungen eingespart werden könnten. Dafür wird die übliche DEA-Methode durch ein spezielles Bootstrapping-Verfahren ergänzt, um eine Bias-Korrektur durchzuführen und Konfidenzintervalle zu ermitteln. Die geschätzten Einsparpotentiale für Treibhausgasemissionen werden mit den Emissionsreduktionszielen der Europäischen Kommission für 2030 verglichen. Der Vergleich zeigt, dass Produktivitätsverbesserungen ein quantitativ wichtiges Element darstellen, wodurch gegebenenfalls eine substanzielle Verringerung der Treibhausgasemissionen der Europäischen Union möglich ist.

Die zweite Studie präsentiert ein stochastisches Modell, welches darauf abzielt, die unentdeckten Ölmengen sowie den Zeitpfad der Ölentdeckungen und der Ölproduktion für eine ölfördernde Region vorherzusagen. Das Modell kombiniert drei Submodelle: (i) ein empirisch fundiertes Produktionsmodell für individuelle Ölfelder, (ii) ein „Successive Sampling Discovery Model“ nach Kaufman et al. (1975) zur Prognose der Feldgrößen, und (iii) ein stochastisches Modell vom Typ Poisson-Prozess zur Vorhersage der Entdeckungszeitpunkte. Das Modell wird für die ölfördernden Regionen Norwegen und dem Golf von Mexiko (letztere weiter untergliedert in „Flachwasser-“ und „Tiefwasser-Region“) geschätzt und evaluiert. Die Ergebnisse zeigen, dass für Norwegen und für die „Flachwasser-Region“ des Golfs von Mexico die prognostizierten Ölentdeckungen etwas zu niedrig sind im Vergleich zu den tatsächlichen Ölentdeckungen, während für die Tiefwasser-Region des Golfs von Mexico die Prognosen zu hoch sind. Ähnliches spiegelt sich in den Prognosen für die Ölproduktion wider.

Die dritte Studie ist eine multivariate Wavelet-Analyse des deutschen Großhandelsstrommarktes, welche die Bestimmungsfaktoren von Strompreisfluktuationen anhand von täglichen Zeitreihen untersucht. Die möglichen Bestimmungsfaktoren sind Kohle- und Gaspreise sowie die Residuallast (die gesamte Netzlast minus die Erzeugung aus Windkraft und Solarenergie). Die multivariate Wavelet-Methode erlaubt eine detaillierte Untersuchung der Beziehungen zwischen den Zeitreihen nach Zeit und Frequenz, wobei ebenso die Abhängigkeiten zwischen den verschiedenen Zeitreihen berücksichtigt werden. Die Ergebnisse zeigen, dass die Residuallast der wesentliche Bestimmungsfaktor über kürzere Perioden ist, während Kohle- und Gaspreise die wesentlichen Bestimmungsfaktoren über längere Perioden sind. Ebenso findet die Studie, dass der Zusammenhang zwischen den Energiepreisen zeitlich veränderlich ist, was mit Ergebnissen aus anderen Studien übereinstimmt (siehe z.B., Sousa et al. (2014); Aguiar-Conraria et al. (2018)).

Contents

List of Figures

List of Tables

1 Introduction

"Simply put, energy is the only truly universal currency, and nothing (from galactic rotations to ephemeral insect lives) can take place without its transformations" (Smil (2022), p. 21).

Since the industrial revolution, there has been an unprecedented rise in human energy consumption across the various sectors of the economy, such as manufacturing, transportation, agriculture and housing, mainly in the form of fossil fuels (coal, oil, natural gas). As their combustion leads to CO_2 emissions, this has triggered the problem of anthropogenic global warming which is caused by CO_2 and other greenhouse gas emissions (IPCC (2023)).

From a resource point of view the above quote from Vaclav Smil qualifies energy as the "master resource" (Michaux (2021), p. 2; Martenson (2023), p. 125). If energy is the master resource, then oil in particular could be considered the "master" among the set of master resources due to its high energy density (both per unit volume and per unit mass), since it is easy to transport and store, and since it can fuel even the largest engines and vehicles (Hall et al. (2009), p. 35; Miller and Sorrell (2014), p. 2).

The increased global access to energy has furthermore allowed the increasingly rapid exploitation of other mineral and ecosystem resources, also with consequential waste generation, which has lead to the ongoing processes of resource depletion and environmental degradation. Among others, these include biodiversity loss (i.e. species extinction), land-system impacts (e.g. deforestation, soil erosion), overfishing, depletion of some large aquifers, aerosol emissions, plastic pollution, and pollution by chemicals with eco-toxicities such as pesticides (see, for example, Steffen et al. (2015); Raugei (2023), p. 1). Although these processes operate primarily on the regional scale, there are interactions among these processes, and if the scope thereof transgresses certain boundaries this will turn into impacts at the global level (Steffen et al. (2015)). In this sense, Rockström et al. (2009) and Steffen et al. (2015) highlight the existence of multiple "planetary boundaries" that are better not exceeded in order to guarantee the stability of the current conditions prevailing on the Earth which can support human societies.

A crucial paradigm of ecological economics is that economic analysis should be centered around the recognition that the economic system is a subsystem of the surrounding ecosystem or of the Earth system (Daly (2007), pp. 39, 41).[1] The Earth system comprises the biosphere (i.e. the natural ecosystems), the "geosphere" (including the lithosphere where the soil and the deposits of minerals and fossil fuels reside), the hydrosphere (the oceans, ice sheets, etc.), and the atmosphere (Hagens and White (2021), p. 135). It is finite, non-growing, and materially closed (Daly (2007), pp. 9-10), though energetically it is open with a constant inflow of solar energy and an outflow of heat radiation. The Earth system supplies the necessary raw materials for economic production, absorbs waste flows, and delivers ecosystem services without which humans and other species could not survive (e.g. pollination, nutrient recycling, climate regulation).

From the fact that the Earth system is finite and non-growing, the concept that there are biophysical limits to the size and scope of human economic activity and its growth

[1] The terminology "Earth system" is adopted from Steffen et al. (2015). Regarding the Earth as a integrated system that is akin to a complex organism is also the viewpoint behind the Gaia hypothesis (Lovelock and Margulis (1974)).

becomes evident (Meadows et al. (1972); Daly (2007)). These pertain to the limited capacity "of the ecosystem to absorb wastes and replenish raw materials in order to sustain the economy [and the population]" (Daly (2007), p. 9). To give an example, the waste stream of CO_2 emissions could in principle be absorbed by biomass production (i.e. CO_2 uptake in plant mass and topsoil), but the rate at which this can happen is limited (though it can be influenced by human efforts, e.g. via afforestation and farming practices).

With regard to nonrenewable resources use, the issue of limits or sustainability is more difficult to define. For example, one can extend the definition of limited capacity from above by saying that nonrenewable resources should not be extracted at rates that "exceed the rate of development of renewable substitutes" (Daly (2007), p. 14). One can extend this definition to allow for the substitution with other nonrenewable substitutes (which then, of course, should be available in quantities so that their extraction rate does not violate the same definition of limited capacity).

Since global primary energy consumption in 2022 was still to 82% based on fossil fuels (Energy Institute (2023), p. 9), this means that, as of now, global civilization is largely a fossil-fuel based civilization. In fact, this is in part related to the four materials that are most indispensable for modern society, which Smil (2022, p. 77) refers to as the "four pillars of modern civilization: cement, steel, plastics, and ammonia". The large-scale production of these materials is crucially dependent on fossil fuels as an energy source or as a petrochemical feedstock. The process of iron ore smelting is fueled by coking coal and natural gas, cement production involves coal or heavy fuel oil, plastics are largely made from oil and natural gas, and ammonia is synthesized from natural gas that is also used as the energy source for the synthesis. In total, the global production of these four materials is responsible for 17% of global primary energy supply, and a quarter of the CO_2 emissions from fossil fuels (Smil (2022), p. 78).

This dependence on fossil fuels for many production processes shows that an important research topic is the study of greenhouse gas (GHG) emission reduction potentials that could arise by improving the productive efficiency of firms, and therefore, of economic sectors. Chapter 3 of this thesis deals with this topic by examining efficiency-related GHG emission reduction potentials by sector (e.g. manufacturing, power generation) for a sample of European countries.

Reconsidering the above paragraph where I have referred to energy and oil in particular as the "master resource", it is hard to imagine the current world without oil. Aside from being used to produce plastics, oil is currently essential for powering heavy transportation, including heavy machinery in agriculture and mining. It is vital for transportation also in other ways: synthetic rubber for tires is made from oil (or natural gas),[2] and the asphalt pavement of roads contains bitumen, a viscous residue from oil distillation.[3] Since oil has provided 31.6% of global primary energy in 2022 (Energy Institute (2023), p. 9), and considering the nonrenewable nature of this fossil resource, this makes it important to study the future prospects of oil availability. Chapter 4 of this thesis considers this topic by formulating a predictive model for oil production, which is then applied to the oil-producing regions of Norway and the U.S. Gulf of Mexico.

Both the resource depletion issue and the issue of global warming highlight the importance

[2]See https://www.britannica.com/science/rubber-chemical-compound

[3]See https://www.britannica.com/science/bitumen

of developing and deploying substitutes for fossil fuel use. The renewable energy flows harnessed by solar panels and wind turbines, hydropower plants, and biomass use are widely regarded as substitutes for fossil fuels, and a massive expansion of these renewable technologies is planned by many of the world's nations. Alternative candidates could be the latest generation nuclear power reactors and geothermal power generation. Since these energy technologies produce mainly electricity (except for biomass), if they are to substitute for the fossil energy that is used in non-electric sectors, their expansion would require the further electrification of the other sectors, by which the electricity sector would further rise in importance.

The analyses presented in Chapters 3 and 4 necessarily employ a long-run perspective since efficiency-related emission reduction potentials require time to be realized and since oil discovery and depletion are per se lengthy processes. Chapter 5 of this thesis considers, among other things, the energy mix of German electricity generation. This is now approached from a short-run perspective by examining how the electricity price is dynamically influenced by the electricity mix and fuel prices. In particular, this study of electricity price formation deals with a situation where electricity is generated from a mix of fossil and renewable energy sources, and where the renewable generation capacity is subject to a steady expansion. The German electricity market is well suited for this because Germany is currently a forerunner where large investments into renewable energy expansion have been made. In 2022, 40% of the total electricity demand was provided by wind and solar, 11% by hydropower and biomass, while 34% was provided by coal and 8% by gas (data obtained from the Bundesnetzagentur). In Chapter 5, therefore, I examine for the German electricity market with a suitable method the relative importance of the "periodic oscillations" in renewable energy generation versus those in fossil energy prices for explaining the "periodic oscillations" in daily wholesale electricity prices.

1.1 Outline of the Thesis

The doctoral thesis is structured as follows. Chapter 2 starts with an empirical examination of the global macroeconomic relationships between primary energy consumption, GHG emissions, and GDP. Afterwards, historical trends of GHG emissions, oil production, and renewable energy generation shares are briefly explored. The chapter is concluded by briefly sketching a framework for macroeconomic production functions with energy as an input and GHG emissions as an undesirable output.

Chapter 3 presents a stochastic nonparametric efficiency analysis in which the emission of greenhouse gases are taken into account as bad outputs. The chapter reports point estimates and confidence intervals for GHG emission reduction potentials for 7 sectors (comprising the largest GHG emitting sectors) in a sample of 16 large European countries. In addition to GHG emissions, the variables used in the efficiency analysis are gross value added as the good output and the conventional labor and capital measures as inputs. The nonparametric approach of the efficiency analysis employs directional distance functions (DDF) to find the maximum possible reduction of the bad outputs which is feasible within a convex technology set that envelopes the input-output combinations of the sample countries. The first variant obtains in this way estimates of the reduction potentials for a single GHG aggregate, computed as the sum of CO_2, CH_4 and N_2O emissions. In other variants the reduction potentials for splits of the different greenhouse gases are examined as well. In a further variant the combination of GHG emission reduction and

output enhancement is examined. In these cases where the inefficiency is assessed in the direction of multiple bad outputs or of a combination of bad and good outputs, the direction vector of the DDF is determined endogenously with the inefficiency. Since the nonparametric efficiency analysis provides only point estimates that are also biased, which is a peculiarity that generally arises for frontier function estimation, the computations are extended by a specifically designed bootstrapping procedure to compute confidence intervals and correct the bias. Thus, the main results of Chapter 3 are bias-corrected GHG emission reduction potentials with confidence intervals. In the final sections of Chapter 3, some policy implications of the analysis are discussed and the magnitudes of the emission reduction potentials are put into perspective by comparison with the emission reduction targets for 2030 as set by the European Commission (see EU (2020)).

Chapter 4 presents a stochastic model for oil production in a region with empirical applications to oil production in Norway and the U.S. part of the Gulf of Mexico. The chapter starts with a literature review of previous mathematical approaches for modeling oil production in a region, such as those based on fitting suitable curves (e.g. bell-shaped) to regional production time series, often called "top-down" models. The chapter then formulates in great detail a stochastic "bottom-up" model for the regional rate of oil production that combines three mathematical submodels. The first one is an empirically-derived field-level production model, where it is assumed that each individual oil field follows a certain deterministic production profile over time, which can vary by field size (i.e. the amount of recoverable oil). The second model is based on certain axioms about how the discovery of new fields proceeds. The main axiom postulates that the fields are not discovered as a random sample but instead via a successive sampling scheme where field size is the key determinant of the discovery order, henceforth called size-biased sampling. The third model is about the dates of "when" the new fields are discovered. Here I use the stochastic process model referred to as a pure birth process. I also introduce time-varying behavior into this model for the discovery times by using extrapolations of certain temporal trend functions, which either pertain to the pace of exploration well drilling or directly to the pace of new field discoveries. A large part of the methods section of Chapter 4 covers the statistical estimation procedure for the size-biased sampling model in detail. The estimation allows to infer from a sample of oil fields (or rather their sizes) a predictive distribution for the undiscovered resource potential (i.e. the remaining amount of oil that "is still out there" in the undiscovered fields). The methods section then proceeds with the pure birth process model for the discovery times, after which the overall properties of the combined model for regional oil production are summarized, and expectations regarding the dynamics of the model are discussed. The empirical part of Chapter 4 first reports the parameter estimates of the size-biased sampling model jointly with the resulting estimates of undiscovered resources. These are then compared to the estimates that are published by the respective official agencies for the regions. The final subsection presents the main results of the chapter, namely the projections of regional oil production (and discovery) for each year beyond the time period that was used to estimate the parameters. These projections are conditional on certain scenarios for future exploration well drilling or for the future pace of new field discoveries. The projections comprise a whole predictive distribution that is shown as a mean forecast with an associated quantile range.

Chapter 5 presents a conditional wavelet analysis that examines the relation between German wholesale electricity prices and a few possible determinants, using daily data from 2015-2023. These determinants are the total electricity demand, wind and solar electricity generation, the residual load (i.e. total demand minus wind and solar generation), and

proxy price series for the import prices of hard coal and natural gas. The chapter starts with a brief review of related literature and continues with the methodology of wavelet and cross-wavelet analysis, where also the novel multivariate wavelet methods (partial and multiple wavelet coherence) which are applied in this chapter are explained. After a brief data description, the empirical results section first considers each time series separately via univariate wavelet analysis. Then, the bivariate wavelet analysis examines the co-movement relations between the electricity price and the possible price determinants in time-frequency space, where the frequency dimension corresponds to cycles with period lengths ranging from 2 days to 2-4 years. The bivariate results are presented graphically via wavelet coherence plots jointly with an indication of the statistical significance of the coherence, which is computed from simulations of independent surrogate series. Also derived from the bivariate analysis are phase difference plots that allow to assess the lag-lead relation between the pair of series and how this relation changes over time. Finally, the analysis in Chapter 5 goes beyond the usual bivariate wavelet analysis by computing also partial and multiple wavelet coherences. The partial wavelet coherence allows to analyze in time-frequency space the relation between the electricity price and the quantity series (e.g. the residual load series) after elimination of the potentially distorting influence of the price series (i.e. coal and gas prices), and vice versa. Besides, the multiple wavelet coherence allows to quantify the explanatory power that the quantity and price series jointly have for the electricity price across the time-frequency space. In the final part of Chapter 5, I summarize the insights gained from the wavelet analysis regarding the question of which of the possible electricity price determinants are most important at which frequency, and how this has changed during 2015-2023.

Chapter 6 briefly reviews the key insights gained from the analyses in Chapters 3 - 5, and based on these insights, possible avenues for future research are outlined.

2 Energy as a Factor of Production

To motivate the studies in Chapters 3 - 5 which deal with GHG emissions, energy (oil) supply and energy (electricity) prices, sect. 2.1 starts with a descriptive examination of the macroeconomic dependence on energy consumption by studying the relations between energy, GDP, and GHG emissions for the economies of the world. Sect. 2.2 proceeds with a further data description of historical trends for GHG emissions, oil production, and renewable energy generation shares. Finally, sect. 2.3 offers a brief sketch of a macroeconomic production function in which the previously established central role of energy and associated GHG emissions can be accounted for explicitly.

2.1 The Macroeconomic Relations between Energy, GHG Emissions, and GDP

The data used in this section are GDP and population data from the Penn World Table 10.01,[4] primary energy consumption data published by the OECD (called "primary energy supply" on their website),[5] and GHG emission data published by the Potsdam Institute for Climate Impact Research (PIK).[6] The PIK emission time series are officially called the PRIMAP-hist national historical emissions (version 2.1) time series, and the sources and methods used for the construction of this dataset are described in detail by Gütschow et al. (2016; 2019). The GHG emission series include all Kyoto gases (excluding land use, land use change and forestry, LULUCF) and are expressed in tons of CO_2 equivalents.

Figure 2.1 shows the bivariate relationships for each pair of the three time series (primary energy consumption, GDP, GHG emissions). The time series are computed as sums over the 110 countries for which data during the period 1971-2019 is available, comprising most of the major economies. On the left-hand side the series are depicted in levels. A linear fit is added as a red line, for which the linear regressions are reported above the panels. On the right-hand side the series are plotted as year-to-year percentage changes and as functions of time. The linear regressions of the red-colored series on the black-colored series are reported above the panels. In Figure 2.2, analogous plots are shown when the series are expressed in per capita terms.

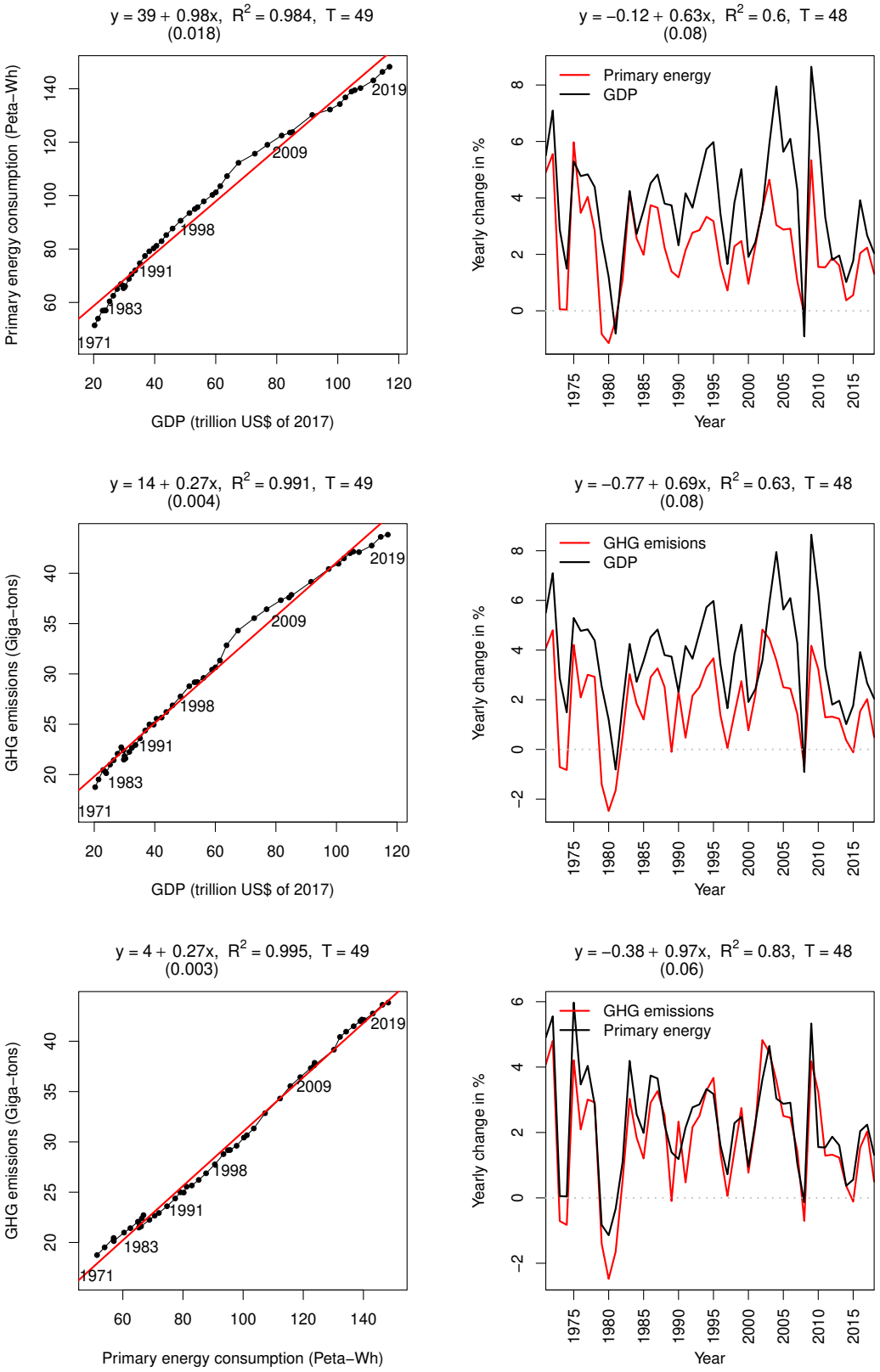
The upper row in Figure 2.1 shows how global energy consumption and GDP have increased over the past 50 years in perfect lockstep. The upper-left panel suggest an almost one-to-one co-movement relation between global GDP and global energy consumption. In order to remove the trend in the time series to avoid spurious correlation between the two nonstationary series, the relation is studied on the right using year-to-year changes. The high correlation is evident and the correlation coefficient amounts to $\sqrt{0.6} \approx 0.77$. The upper-right panel also shows that there were two recessions in terms of global GDP which followed or co-occurred with a decline energy consumption: the first one was in 1981 following a decline in energy consumption in 1979-81 after the second oil crisis, the second one was in 2008 after the 2007–08 financial crisis, which co-occurred with a tiny decline of global energy consumption in 2008.

[4]The data are accessible at https://www.rug.nl/ggdc/productivity/pwt/

[5]The data are accessible at https://data.oecd.org/energy.htm

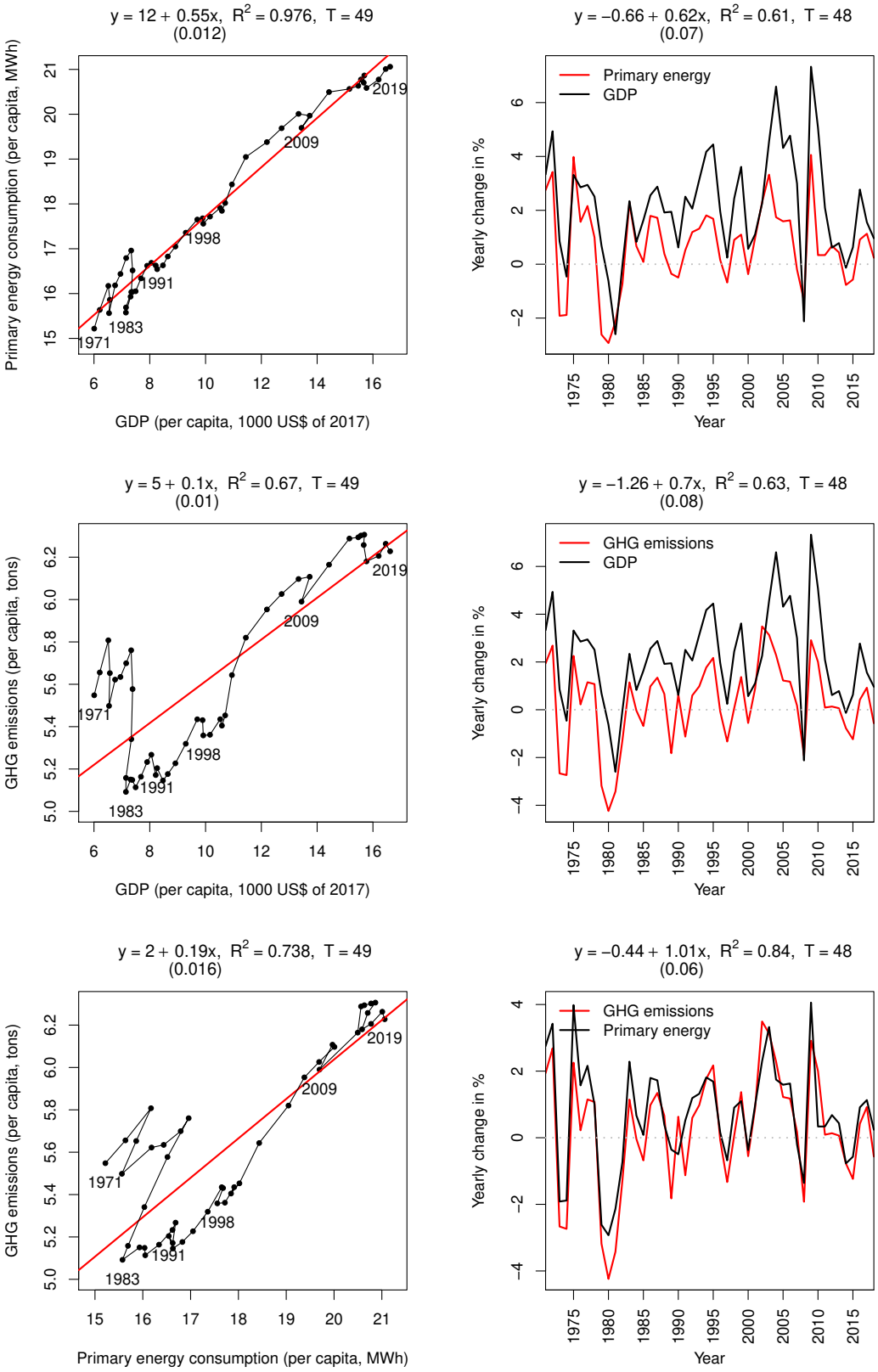
[6]The data are accessible at https://www.climatewatchdata.org/

Figure 2.1: Global time-series relations, in total, 1971-2019



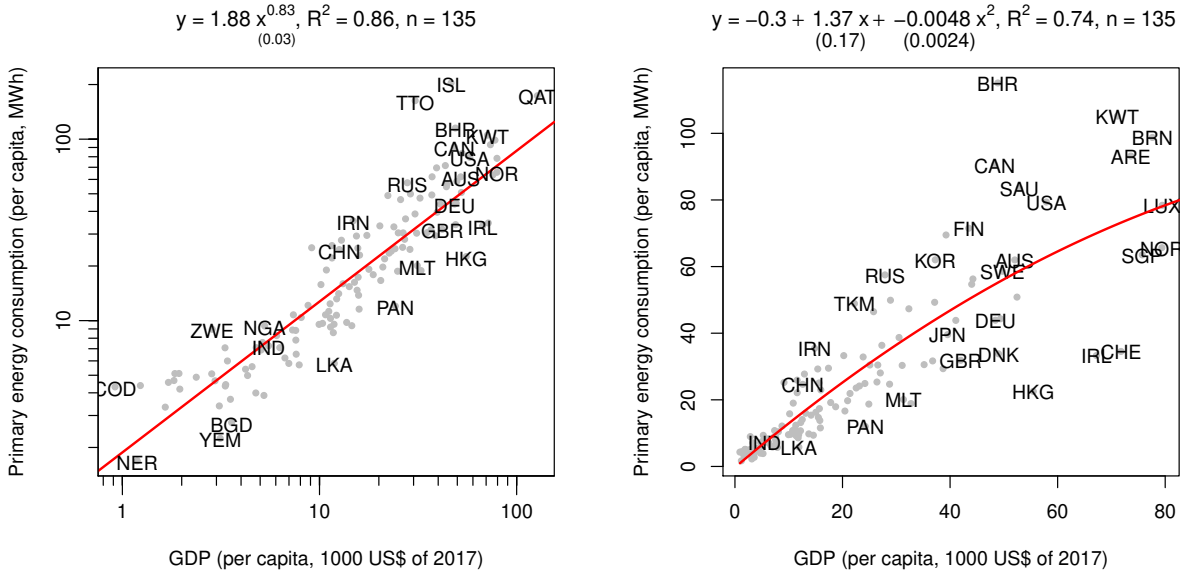
Source: OECD energy data, PIK historical GHG emission data, Penn World Table 10.01.
Note: World totals are computed from 110 countries, comprising all major economies. The equations above the panels on the right show the results from regressing the red time series on the black time series.

Figure 2.2: Global time-series relations, per capita, 1971-2019



Source: OECD energy data, PIK historical GHG emission data, Penn World Table 10.01.
Note: World totals are computed from 110 countries, comprising all major economies. The equations above the panels on the right show the results from regressing the red time series on the black time series.

Figure 2.3: Cross-sectional relation between energy consumption and GDP in 2010-19



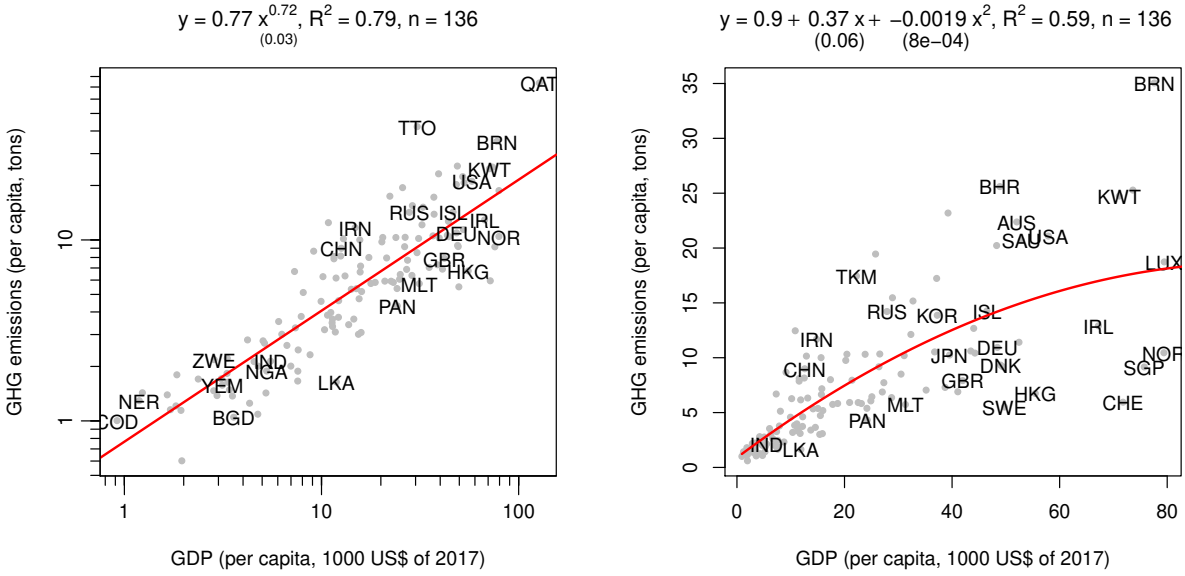
Source: OECD energy data, Penn World Table 10.01.
Note: The variables are taken as medians over 2010-2019. ISO alpha-3 country codes are shown for selected countries. Trinidad and Tobago (TTO), Qatar (QAT) and Iceland (ISL) (not shown in the right panel) are not used to compute the regressions in both panels as they are clearly outliers.

The upper row in Figure 2.2 shows that the relations between per capita energy consumption and GDP exhibit correlations of the same high magnitude. From the upper-right panel showing the short-run relationship one can see that per capita GDP has declined in 1974 after a decline in per capita energy consumption in 1973-74 (the years of the first oil crisis), and also in 2014 jointly with a decline in per capita energy consumption in 2014-15. This brief descriptive analysis suggests that the direction of causation between energy and GDP goes in both ways, as there were clearly recessions that were preceded by a decline in energy consumption, but there were also recessions where the decline in energy consumption occurred only in the same year or afterwards.

Moreover, the bottom row of Figure 2.1 shows a very close association between GHG emissions and primary energy consumption, which according to Figure 2.2, however, is less close when expressed in per capita terms. In particular, the bottom-left panel of Figure 2.2 reveals that the emission intensity of energy consumption has substantially declined after the second oil crisis, but then has increased back again during the 1990s and 2000s.

In Figure 2.3, the relation between per capita energy consumption and GDP is analyzed from a cross-sectional perspective. For a global cross-section of 135 countries, where the variables are taken as the medians over 2010-19, the left panel shows the results from regressing energy consumption on GDP on a double-logarithmic scale, while the right panel shows the results for a quadratic specification. Figure 2.4 shows the same kind of plots and regressions where GHG emissions are used instead of energy consumption. Note that in the left panels the axes have a log-scale. The regression on the left in Figure 2.3 shows that the power function $\text{Energy/capita} = 1.88(\text{GDP/capita})^{0.83}$ fits the macroeconomic data well. As noticed by West (2017) and by Hagens and White (2021, p. 198), such a power law relation is quite similar to a well-known scaling law that applies

Figure 2.4: Cross-sectional relation between GHG emissions and GDP in 2010-19



Source: PIK historical GHG emission data, Penn World Table 10.01.
Note: The variables are taken as medians over 2010-2019. ISO alpha-3 country codes are shown for selected countries. Trinidad and Tobago (TTO) and Qatar (QAT) (not shown in the right panel) are not used to compute the regressions in both panels as they are clearly outliers.

to biological organisms, namely Kleiber's law. This refers to the empirical relationship between the metabolic rate and the mass of animals, which is found to be well-described by a power function with an exponent of 0.75 (West (2017); Thommen et al. (2019)).

In the environmental economics literature, quadratic or cubic functional forms have been employed for the relationship between environmental pollution and GDP, specifically in the literature that studies the so-called Environmental Kuznets Curve (EKC). The EKC refers to an inverse-U-shaped relation between a measure of economic development or output on the x-axis and an environmentally relevant variable on the y-axis. The concept was first studied by Grossman and Krueger (1991; 1995) and Panayotou (1993) with respect to air pollution such as SO_2 emissions, but there were also early studies using CO_2 emissions (e.g. Holtz-Eakin and Selden (1995)). In the following decades a long list of studies has been published on this topic, as recently surveyed by Shahbaz and Sinha (2019). Some of the papers also examine an EKC with energy consumption as the environmental variable, for example Richmond and Kaufmann (2006). From the right panels in Figures 2.3-2.4 it is possible to assess whether the relations between energy consumption / GHG emissions and GDP follows an EKC. Figure 2.3 shows that, although the quadratic term is statistically significant, the resulting curve is not inversely U-shaped since the hypothetical turning point for GDP per capita lies far outside the sample range (at about 143,000 US$ per capita). The picture for GHG emissions in Figure 2.4 is similar, as expected because both variables are strongly tied due to the CO_2 emissions from fossil fuels. The hypothetical turning point is here comparatively lower at about 97,000 US$ per capita. Since this point is still quite outside of the sample range, I conclude that there is also no EKC for GHG emissions and GDP.

In sum, Figures 2.1-2.4 elucidate the very strong connection between energy consumption / GHG emissions and GDP both from a global time-series perspective and from a

country-level cross-sectional perspective. This result is not very surprising from a physical understanding of the role of energy for economic production, i.e. from the point of view of thermodynamics (as expressed for example in the introductory quote to Chapter 1 from Vaclav Smil). As also stated by Ayres and Warr (2009, p. xviii): "In contrast to the neoclassical economic model, the real economic system depends on physical material and energy inputs, as well as labour and capital. The real economic system can be viewed as a complex process that converts raw materials (and energy) into useful materials and final services. Evidently materials and energy do play a central role in this model of economic growth." Besides, Hall et al. (2001, p. 663) point out that "the two laws [of thermodynamics] say that nothing happens in the world without energy conversion and entropy production, with the consequence that every process of biotic and industrial production requires the input of energy". In particular, the second law of thermodynamics implies that "the valuable part of energy (exergy) is transformed into useless heat at the temperature of the environment (anergy), and usually matter is dissipated, too. This results in pollution and, eventually, exhaustion of the higher grade resources of fossil fuels and raw materials" (Hall et al. (2001), p. 664). One of the first economists who stressed the importance of taking into account the relation between the laws of thermodynamics and the economic process was Nicholas Georgescu-Roegen (1971). Among others, these scientists have criticized the prevalent (macro-)economic doctrine where the important role of energy is largely ignored. As Smil (2022, p. 21) states: "Given all of these readily verifiable realities, it is hard to understand why modern economics, that body of explanations and precepts whose practitioners exercise more influence on public policy than any other experts, has largely ignored energy".

2.2 Historical Trends for GHG Emissions, Oil Production, and Renewable Energy

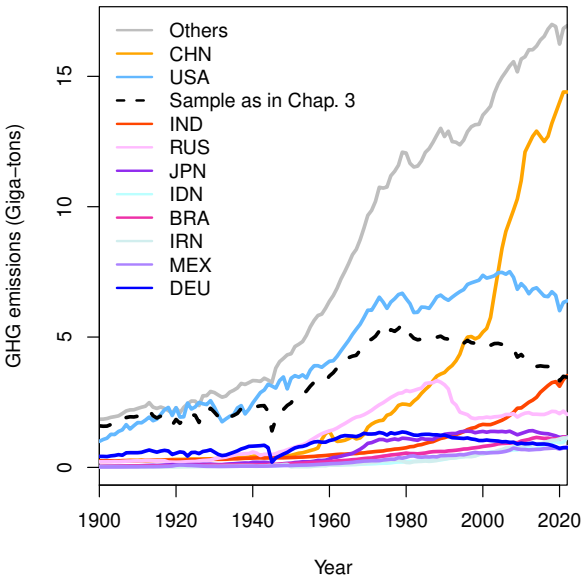
In this section, further historical time series on GHG emissions, oil production, and renewable energy shares are depicted and briefly analyzed.

Figure 2.5 depicts the GHG emissions (measured in tons of CO_2 equivalents) during the time period 1900-2022 for the 10 countries that are the largest emitters as of 2022. As the gray line the figure also shows the aggregate emissions of the remaining countries. To allow for comparison with the sample of 16 European countries examined in Chapter 3, I also added a dashed black line that shows their total GHG emissions. First, one can see that the GHG emissions of the sample from Chapter 3 and of India have been converging so that both are of the same magnitude in 2022. The figure also illustrates the proportions of the GHG emissions of the two largest economies, the United States and China, versus those of "Others". Particularly noteworthy is the unprecedented rise in the GHG emissions of China after 2000. On the other hand, while the Chinese population is more than four times as large as the US population, the Chinese GHG emissions were only about twice as large as the US emissions in 2022.

Figure 2.6 shows the monthly production of liquid fuels for the world (left panel) and for the US only (right panel), obtained from the US Energy Information Administration (EIA). The blue area represents crude oil production and the lightest orange represents natural gas liquids (NGL) production.[7] The NGL category includes the hydrocarbon

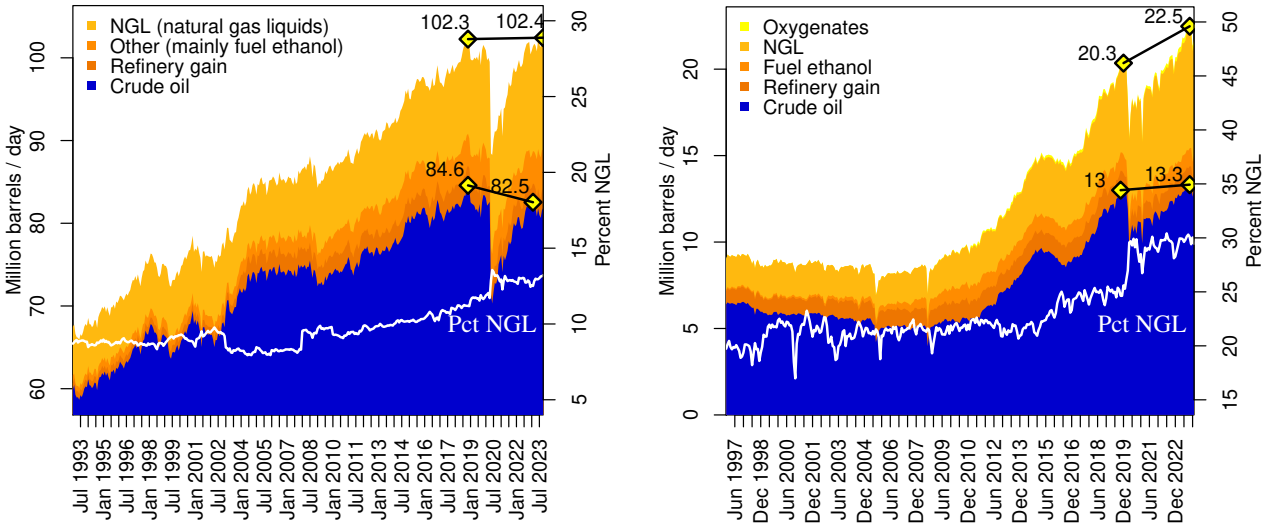
[7]The other orange areas pertain to refinery or processing gain and fuel ethanol production, for defini-

Figure 2.5: GHG emissions of top 10 emitters and others, 1900-2022



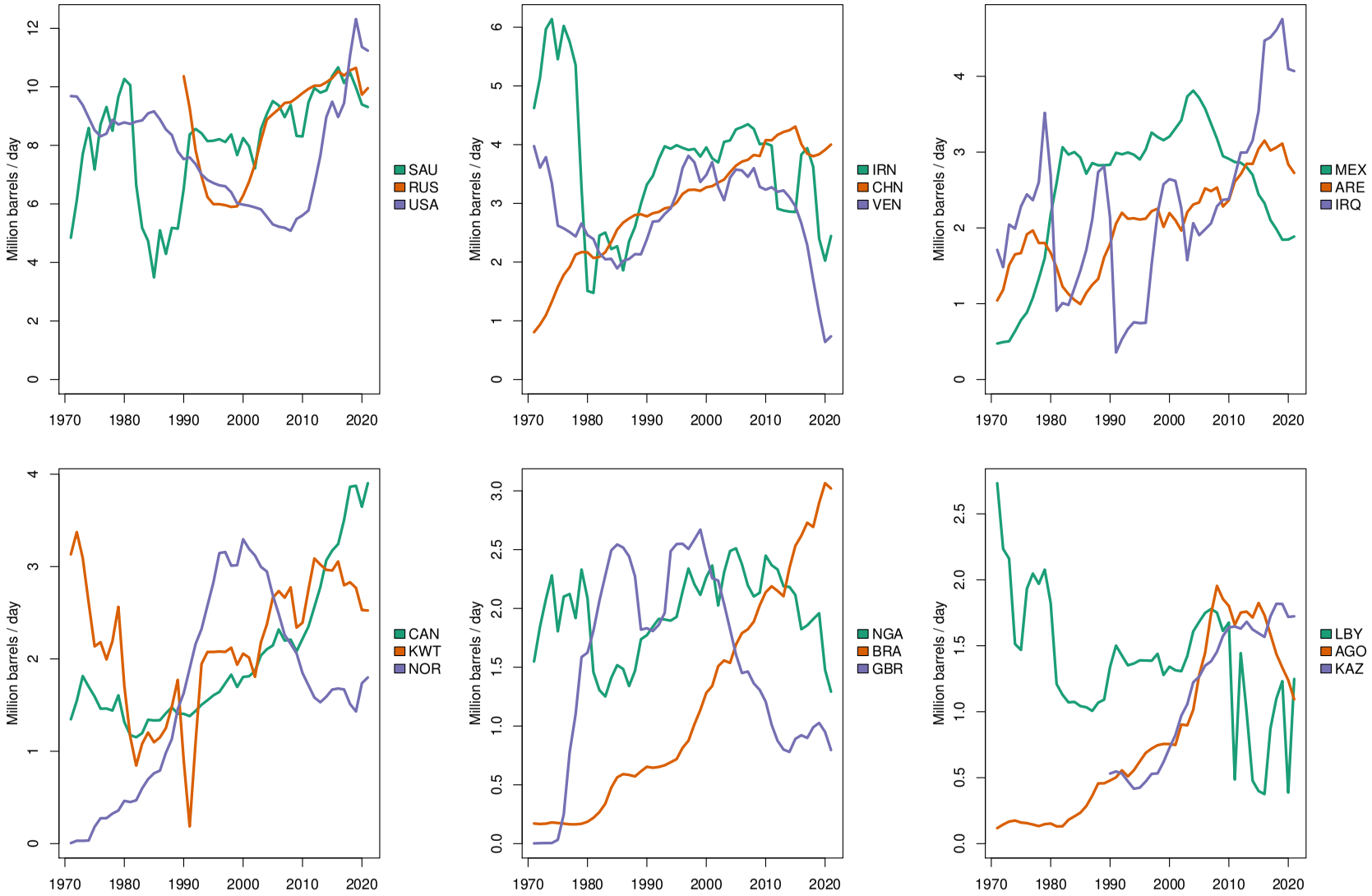
Source: PIK historical GHG emission data

Figure 2.6: World (left panel) and US (right panel) liquid fuel production



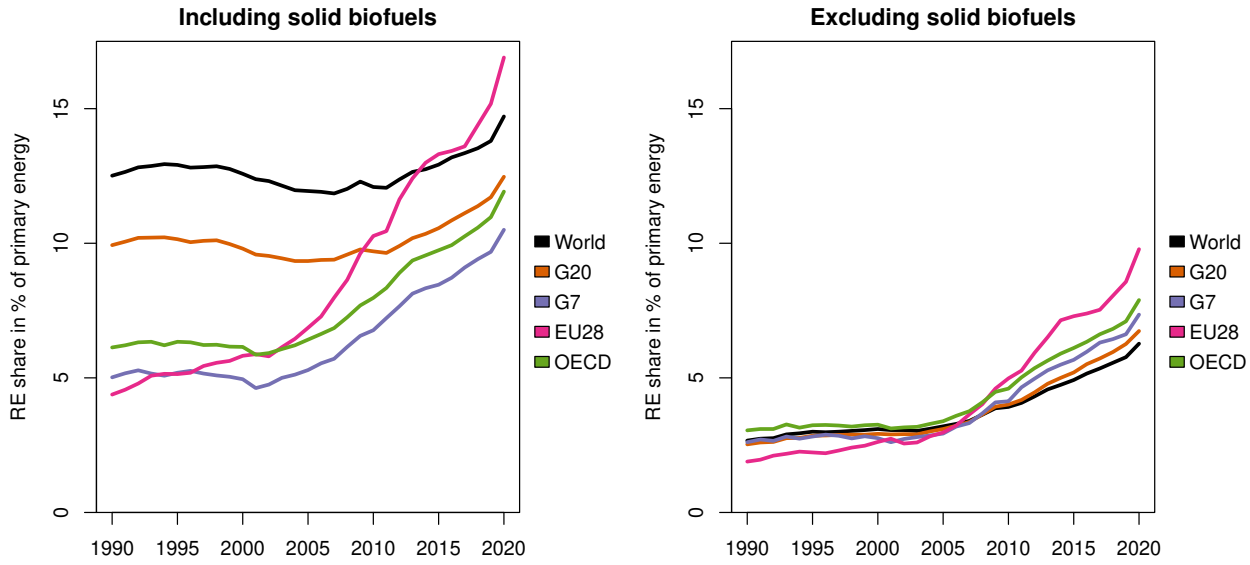
Source: Energy Information Administration (EIA). Own illustration, inspired by similar charts from Art Berman, Labyrinth Consulting Services, Inc.

Figure 2.7: Crude oil production of top 18 oil-producing countries, 1971-2021



Source: OECD energy data on crude oil production

Figure 2.8: Renewable energy (RE) shares in primary energy consumption, 1990-2020



Source: OECD energy data on renewable energy.

molecules ethane, propane, butane, and pentane. The shorter molecules (ethane, propane) are rarely used for transportation but rather for petrochemicals or heating and have a considerably lower energy content than crude oil. Ethane occupies the largest share of NGL production and is almost exclusively used to make plastics.[8] The left panel of Figure 2.6 shows that the recovery of world liquid fuel production after the onset of the COVID crisis has to a larger extent come from NGLs than from crude oil. In fact, global crude oil production so far has not recovered to the peak production that occurred in November 2018, which was more than one year before the beginning of the COVID crisis. Of course, this may still be the aftereffects of the COVID crisis combined with a lagged recovery for crude oil in a capital-intensive industry, but this recent trend highlights the importance of examining the prospects of future crude oil availability. To provide a country-level resolution, Figure 2.7 shows the yearly crude oil production time series for 18 of the largest oil-producing countries over the period 1971-2021, obtained from the OECD website.[9]

Finally, Figure 2.8 shows the shares of renewable energy in primary energy consumption for several country groups, obtained from the OECD website.[10] One can see a shift of the curves when comparing the left panel (with solid biofuels) to the right panel (without solid biofuels), but the rates of increase after 2005-2010 are similar in both panels, showing that this recent increase has come from the new renewables (solar, wind, non-solid biofuels) or hydropower. Still, when excluding solid biofuels, the renewable energy shares in 2020 amounted only to 6.3% for the world and to 9.8% for the EU-28. Overall the EU countries show the fastest rise of the renewable energy share.

tions see https://www.eia.gov/tools/glossary

[8] See https://www.eia.gov/todayinenergy/detail.php?id=5930

[9] See 'crude oil production' at https://data.oecd.org/energy.htm

[10] See 'renewable energy' at https://data.oecd.org/energy.htm

2.3 A Framework for Macroeconomic Production Functions with Energy and GHG Emissions

To conclude this chapter, what follows is a brief sketch of a framework for macroeconomic production functions in which the central role of energy consumption and of associated GHG emissions, as consistent with the conclusions from sect. 2.1, is explicitly taken into account. Specifically, I consider a function that maps the three inputs capital (K), labor (L), and energy consumption (E) into the economic output such as GDP (Y), jointly with a second production function for GHG emissions (U) as a by-product from producing the economic output. Besides, energy consumption is further subdivided into fossil fuel energy (E_f) and energy from other sources (E_o), where $E = E_f + E_o$. The general framework thus looks as follows:

$$\begin{aligned} Y &= F\big(KA_K^Y, LA_L^Y, (E_f + E_o)A_E^Y\big), \\ U &= G\big(K/A_K^U, L/A_L^U, E_f/A_E^U\big), \end{aligned} \tag{2.1}$$

where also the partial derivatives of the functions F and G with respect to each argument are nonnegative.

The function F contains as parameters the augmentation indices of capital (A_K^Y), labor (A_L^Y), and energy (A_E^Y) which can be interpreted as reflecting changes in technology or efficiency that improve the capability of the respective inputs to produce GDP. For the function G, the parameters A_K^U, A_L^U, and A_E^U are analogous augmentation indices that reduce the GHG emissions per input use. To give an example, an increase in A_E^U (lowering GHG emissions per fossil energy consumption) could occur due to the substitution of coal with gas (since gas contains less carbon per unit of energy), due to reduced leakage of methane emissions at gas or oil production sites, or due to the application of carbon capture and storage (CCS).

The next chapter now continues with an environmental efficiency analysis of 7 main GHG emitting sectors in 16 European countries, from which GHG emission reduction potentials are estimated for the sectors and countries. While the analysis does not explicitly include energy consumption, an implicit estimation of the functional relation (or rather of the efficient frontier) between the other variables (GHG emissions, economic output, labor and capital) is conducted by using a nonparameteric efficiency analysis approach.

3 Greenhouse Gas Emission Reduction Potentials in Europe by Sector: A Bootstrap-Based Nonparametric Efficiency Analysis[11]

3.1 Introduction

The reduction of greenhouse gas (GHG) emissions on a global level is the key measure to counteract the detrimental effects of climate change and global warming. Other approaches to limit global warming, like carbon removal or geoengineering approaches are either infeasible or extremely risky (see Nordhaus (2019, p. 1998) for a clear statement). This is largely undisputed in the economic literature (see the survey articles by Myhre et al. (2001), Aldy et al. (2010), Hsiang and Kopp (2018) and Tol (2018), among others) and is the basis for several international agreements. The most prominent agreements are the Kyoto Protocol of 1997 and the Paris Agreement on climate change of 2015 to reduce GHG emissions to reach the 2°C target, meaning the stabilization of the increase in temperature at "well below 2°C above pre-industrial levels".[12]

The European Union (EU) as a key actor in this area has achieved an agreement among its member countries to reduce GHG emissions by 40% until 2030, 60% until 2040 and 80% until 2050, compared to the levels of 1990 (EU (2011), p. 3). Recently these targets have been tightened to reduce GHG emissions by 55% until 2030, 80% percent until 2040 and to reach climate neutrality by 2050, also accounting for the effects of carbon removal technologies, land use change and forestation (see EU (2020) and especially figure 1 therein).

Efforts to improve the productive efficiency of sectors could be a potentially important building block of an emission reduction strategy. Therefore it is important to know to which extent GHG emissions could be reduced by achieving productive efficiency while holding the economic inputs and outputs constant. In our companion paper Krüger and Tarach (2020) we applied nonparametric methods of efficiency analysis in the presence of undesirable outputs derived from a variant of data envelopment analysis (DEA) to give an account of the potential reductions of GHG by country and sector for the period 2008-2016. The main finding is that efficiency improvements can contribute considerably to emission reduction, albeit the extent to which the measured potentials could be realized in practice remains open. However, the measurement approach used in the companion paper is purely deterministic and prone to biases. Furthermore, no account of the estimation uncertainty is provided there.

In this paper we pick-up these issues by combining the nonparametric efficiency measurement approach with a specifically designed bootstrapping procedure to achieve a bias correction and to compute confidence intervals for assessing estimation uncertainty. To our knowledge this is the first time that a nonparametric approach combined with stochastic elements is applied in an environmental efficiency measurement context. We report estimates of aggregate emission reduction potentials for 16 major EU countries and 7 main

[11] This chapter is joint work with Jens Krüger and is published in *Environmental and Resource Economics*, see Krüger and Tarach (2022).

[12] This 2°C target is defined in Article 2 of the Paris Agreement jointly with the plea to pursue an even tighter target of 1.5°C, see https://newsroom.unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.

sectors of the private economy. As emissions we consider a broad GHG aggregate as well as splits to single GHG (CO_2, CH_4 and N_2O). The results show that the bias correction leads to larger emission reductions compared to the "raw" measures from our companion paper which are based on the purely deterministic approach. We can show that the potentials for emission reduction are concentrated in certain countries and sectors. In addition, we find that the estimation uncertainty is substantial in these cases.

In contrast to much of the literature on eco-efficiency which is also concerned with emission reduction on a macroeconomic level or major sectors we assess the contribution of potential efficiency improvements to the EU reduction targets by expressing them as potential reductions measured in physical units, i.e. CO_2 equivalents (CO_2e). The usual habit in the literature (see Camarero et al. (2014), Färe et al. (2004), Korhonen and Luptacik (2004), Kortelainen (2008), Kuosmanen and Kortelainen (2005), Rashidi and Farzipoor Saen (2015), Zaim and Taskin (2000), Zhou and Ang (2008) and Zofío and Prieto (2001), among others) is to focus on relative measures instead. More closely related to our analysis are studies such as Domazlicky and Weber (2004) and Krautzberger and Wetzel (2012) which are also based on a methodological setting employing directional distance functions and are also confined to specific industries.

The exposition in this paper starts with a description of the data and the country-sector coverage in section 3.2. This is followed by the description of the nonparametric methodology we use to obtain our estimates of emission reduction potentials in section 3.3. In this section, the implementation of the bootstrapping approach as well as the computation of the bias-corrected measures and the confidence intervals are also outlined. Section 3.4 contains the discussion of the results from several specifications of the undesirable outputs. The specifications comprise a single total GHG aggregate as well as splits to CO_2, CH_4 and N_2O. We also discuss the results of a variant where possible enhancements of the economic output are permitted in addition to the emission reductions. Policy recommendations are provided at the end of the section. The final section 3.5 concludes with an evaluation of the contribution of the emission reduction by efficiency improvements to the EU emission reduction targets and discusses the feasibility of the potential reductions measured.

3.2 Data Description

The data required for the efficiency analysis comprise the inputs, the good (desirable) outputs and the bad (undesirable) outputs, i.e. the emissions of greenhouse gases. In the subsequent measurement of inefficiency and the potential emission reduction derived from the inefficiency measure we always include the two conventional inputs labor and capital as well as value added as the single economic output. The emissions as undesirable outputs are used in different forms. As the description of the methods will show, the inefficiency is measured as the potentially reachable *enhancement* of the good output and/or the potentially reachable *reduction* of the emissions. The economic data, meaning the inputs and the good (desirable) output are taken from the EU-KLEMS database. The November 2019 release we use is described by Stehrer et al. (2019) and can be obtained from https://euklems.eu. Labor input is measured in total hours worked by employees (comprising self-employed persons and expressed in full-time equivalents). Capital input is quantified by the real fixed capital stock (at constant 2010 prices). The output variable

is gross value added (also at constant 2010 prices).[13] Using this variable is associated with a much more comprehensive data coverage compared to the alternative of using a gross output measure with materials and energy as additional input variables.[14]

The emissions data to quantify the bad (undesirable) outputs[15] are taken from two sources.[16] As greenhouse gas (GHG) emissions, we focus on the three main greenhouse gases (GHGs) which are emitted by anthropogenic sources, namely carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O). The global warming potentials usually differ for each GHG, but they can be converted to CO_2 equivalents (abbreviated CO_2e and measured in tons, kilotons or megatons). CO_2 emissions are retrieved from the World Input Output Database (WIOD) described in Timmer et al. (2015) and can be downloaded from http://www.wiod.org. The data for CH_4 and N_2O emissions are retrieved from the Eurostat Air Emission Accounts (AEA).[17] In the AEA database, CH_4 and N_2O emissions are already expressed in tons of CO_2e and so we obtain our measure of total GHG emissions by simply adding them to the CO_2 emissions from the WIOD. There are further GHGs which are of minor quantitative importance and therefore neglected.[18]

All three major GHGs have specific anthropogenic sources. CO_2 emissions stem primarily from burning fossil fuels (coal, oil and natural gas), but also from industrial processes such as the manufacturing of cement. In addition, CO_2 is emitted from land use, land use change and forestry (LULUCF). Although its global warming potential per ton is less than that of CH_4 or N_2O, CO_2 is quantitatively the most important GHG. In 2010 CO_2 emissions (without LULUCF) accounted for 82% of total GHG emitted by the EU (Debelke and Vis (2015), p. 96).

CH_4 has an atmospheric lifetime of 12 years, meaning that on average it stays in the atmosphere for only 12 years before it is broken down into CO_2 and water (Hsiang and Kopp (2018), p. 12). It has a global warming potential of 25 CO_2e (meaning one ton of CH_4 has the global warming potential of 25 tons of CO_2, Eurostat (2015), p. 105). The two major anthropogenic sources of CH_4 emissions are industrial livestock farming and the exploitation of fossil fuels. Natural gas (largely consisting of CH_4) may be leaking when recovered from gas or oil fields or during transport and storage. CH_4 is also contained in coal beds (coal mine methane), especially in deeper coal beds and coals with higher carbon content (i.e. hard coal), and may similarly leak during coal mining (Kholod et al. (2020)). In the EU, CH_4 emissions already declined between 1990-2010 by 32% (Debelke and Vis (2015), p. 96).

N_2O is a very potent GHG with the same global warming potential as 298 tons of CO_2 (Eurostat (2015), p. 105) during an atmospheric lifetime of 116 ± 9 years (Tian et al.

[13] We always mean the good (desirable) economic output when we simply refer to the output in the following.

[14] This alternative would also increase the dimensionality of the input-output space which is a crucial issue for nonparametric analyses in general.

[15] We subsequently refer to emissions when we mean the bad (undesirable) outputs.

[16] These data bases are used instead of the Emissions Database for Global Atmospheric Research (EDGAR) because of their conformability to an economic sector classification and their coverage of more recent periods.

[17] These data can be accessed at https://ec.europa.eu/eurostat/web/products-datasets/-/env_ac_ainah_r2.

[18] Further anthropogenic GHGs are sulphur hexafloriode, hydrofluorcarbons and perfluorcarbons, which are not included in our measure of total GHG emissions. They made up only 2% of total GHG emissions in the EU-28 in 2010 (Debelke and Vis (2015), p. 96), slightly rising to about 2.5% in 2018 (EEA data).

(2020)). In addition, N_2O has a depleting effect on the stratospheric ozone layer. The major anthropogenic source of N_2O is the agricultural sector, in particular the large-scale use of nitrogen fertilizers. According to Tian et al. (2020) agricultural emissions accounted for about 70% of anthropogenic N_2O emissions globally in 2007-16. Other comparatively smaller anthropogenic sources include the fossil fuel and chemical industry. In contrast to rising or stagnant N_2O emissions in most other countries globally, European emissions from agriculture declined by 21% between 1990-2010 (Tian et al. (2020), p. 254), which the authors attribute to European agricultural policies favoring more efficient fertilizer use. Besides, non-agricultural N_2O emissions in the EU were reduced even more strongly during that period, mainly due to improved abatement technologies in the chemical industry (Tian et al. (2020), pp. 253-255).

Assessing the data coverage in the database we are able to achieve an almost complete coverage for 16 countries and 7 sectors during the period 2008-2016 on a classification of sectors (industries) according to NACE Rev. 2 (equivalent to ISIC Rev. 4). The countries covered comprise (with World Bank country codes in parentheses):

Austria (AUT)	Germany (DEU)	Poland (POL)
Belgium (BEL)	Greece (GRC)	Slovakia (SVK)
Czech Republic (CZE)	Ireland (IRL)	Spain (ESP)
Denmark (DNK)	Italy (ITA)	Sweden (SWE)
Finland (FIN)	Netherlands (NLD)	United Kingdom (GBR)
France (FRA)		

The sectors covered are:

A Agriculture, Forestry and Fishing
B Mining and Quarrying
C Manufacturing
D Electricity, Gas, Steam and Air Conditioning Supply
E Water Supply Sewerage, Waste Management and Remediation Activities
F Construction
G Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles
H Transportation and Storage

The emissions data in the AEA database are only available for a sector combining the sectors D and E. So we had to aggregate the economic input-output data of the sectors D and E to a combined sector, henceforth named DE. Cross checking assures that the sums of the values of the sectors D and E are very close to the values of the combined sector DE which is also available in the EU-KLEMS data.[19] Since the sector D is considerably larger than E in most countries we refer to the combined sector DE frequently as "energy" or as "energy and water" in the subsequent discussion.

We exclude Estland, Lithuania, Luxembourg and Slovenia from our analysis despite full data coverage. The reason is that these are very small countries and Luxembourg is merely a large city rather than a country. Including those small countries can severely bias the entire efficiency analysis when they determine parts of the frontier function and overstate the potential emission reductions. Growiec (2012) provides further discussion

[19] An exception are two capital stock values of Belgium in 2008 and 2009 where the sums of the values of the sectors D and E deviate from those of the combined sector DE by 18 and 5 percent, respectively. Here we use the time series of the sum of the single sectors which looks more plausible than the time series of the combined sector. In the case of Spain only data for the combined sector are available and therefore these data are used directly.

of this issue. In some of these countries we also suspect recording errors in the data for some sectors (e.g. zero emissions in sector G in Slovenia).

The value added and capital stock data are directly expressed in Euro for the majority of the countries (appropriately deflated with base year 2010). In the case of the non-Euro countries Czech Republic, Denmark, Poland, Sweden and the United Kingdom these variables are expressed in the respective national currencies. To convert the data to a common currency we use purchasing power parities (PPPs) from the OECD National Accounts Statistics (OECD (2020)). While exchange rates only convert currencies, PPPs also take account of different price levels of the countries. This is important since price levels tend to be systematically higher in high-income countries than in low-income countries. Using exchange rates would therefore overstate the values of the variables in the case of high-income countries and understate them in low-income countries. Instead, PPPs convert expenditures to a common price level. This is also important for countries with a common currency (as the Euro) which also can have rather different national price levels.[20]

We split these data in two five-year subperiods $t_1 = 2008\text{-}2012$ and $t_2 = 2012\text{-}2016$ and take medians over these subperiods for the subsequent empirical analysis. This eliminates the effects of single or even two outlying observations and makes the efficiency analysis more robust. The way of taking medians to robustify the analysis is in our view preferable to the alternative of outlier detection by methods such as those proposed by Wilson (1993) and subsequent outlier elimination. This procedure also solves the problem with two missing values in sector C of Ireland.[21] Thus, when we refer to the first and second subperiod in the following we always mean the medians of the inputs and outputs (including emissions) over the indicated five-year intervals.

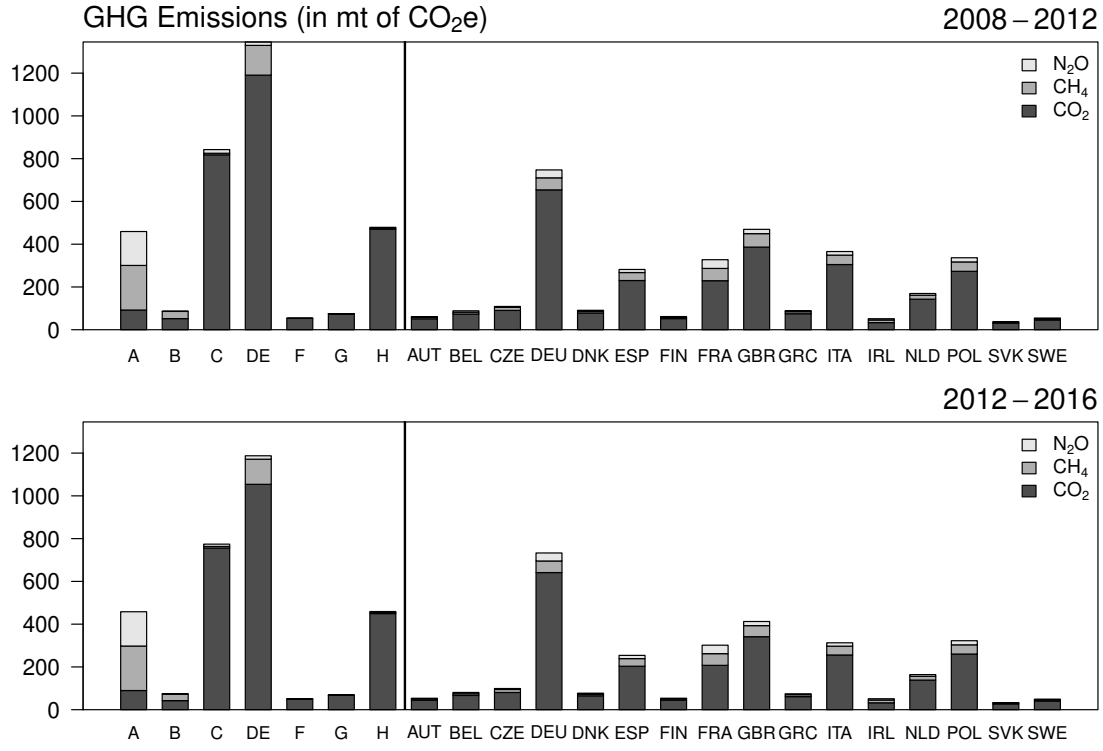
The aggregate GHG emissions over all countries and sectors are 3341 mt of CO_2e in the first subperiod, declining to 3070 mt in the second subperiod. Figure 3.1 shows stacked barplots of the three GHG emission variables for both subperiods (the corresponding data are reported in Table A1 in the appendix). The left-hand side of each plot depicts the bars for the sectors, followed by the bars of the countries on the right-hand side (separated by the thick vertical line). This kind of plot gives a succinct summary of the distribution of the aggregate emissions over sectors and countries jointly with an indication of the distribution of the different GHGs (CO_2, CH_4 and N_2O in mt of CO_2e). More descriptive information on the data is discussed in the companion paper of Krüger and Tarach (2020).

From Figure 3.1 we immediately see that the sectors C and DE are most emission intensive, while A and H also contribute considerably, and the remaining sectors (B, F and G) are of minor importance. CO_2 is the quantitatively most important emission category in all sectors except A where CH_4 and N_2O emissions are dominating. CO_2 is the main emission category in all countries, including those with large aggregate emissions (Germany, Spain, France, the United Kingdom, Italy and Poland), although the contribution of CH_4 and N_2O is also visible here. While the overall quantity declines from the first to the second

[20] PPPs are also central for the construction of comparable national accounts provided in the Penn World Table (see Feenstra et al. (2015)).

[21] In the case of Ireland the capital stock values for the final years 2015 and 2016 are missing in sector C. Since the preceding values 2012-2014 show a rising trend (and capital is an accumulating stock variable) we can safely suppose that the missing values are larger than the value in 2014. Then taking the 5-year median over the subperiod 2012-2016 will result in just the value of 2014 irrespective of the exact magnitudes of the missing values.

Figure 3.1: GHG emissions across sectors and countries



subperiod, the distribution of the emissions across sectors and countries is rather similar in both subperiods.

3.3 Nonparametric Efficiency Measurement and Bootstrapping

For the estimation of the potential emission reductions we apply nonparametric methods of efficiency analysis. These methods are an extension of data envelopment analysis (DEA), developed by Charnes et al. (1978) and Banker et al. (1984). The specific modification we rely on is based on the device of the directional distance function (DDF), introduced by Chambers et al. (1996) and extended to an environmental context by Chung et al. (1997). This approach allows to measure inefficiency as the distance to a piece-wise linear frontier function along a mix of possible reduction of inputs and enhancement of some outputs (the good, desirable outputs), while other outputs (the bad, undesirable outputs) are supposed to be reduced (see Färe and Grosskopf (2004)). This property of reducing outputs allows to incorporate undesirable outputs like GHG emissions in a consistent way (Zhou et al. (2008b)). Like in DEA, here also no price information is required and no functional form assumptions about the underlying technology (e.g. a production function) need to be imposed. These are major advantages of the nonparametric approach.

3.3.1 Technology Set

The nonparametric approach of efficiency analysis is based on the concept of an abstract technology set, comprising the feasible input-output combinations. It can be stated as

$$\mathcal{T} = \{(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}) \in \mathbb{R}_+^{m+s+r} : \boldsymbol{x} \geq \boldsymbol{0} \text{ can produce } (\boldsymbol{y}, \boldsymbol{u}) \geq \boldsymbol{0}\}, \tag{3.1}$$

where $\boldsymbol{x}$ denotes the m-vector of the input quantities, $\boldsymbol{y}$ the s-vector of the quantities of the good (desirable) outputs and $\boldsymbol{u}$ the r-vector of the quantities of the bad (undesirable) outputs.[22] Since we are dealing with sectors within countries it is suitable to suppose that each sector operates with a different technology set.[23]

To impose some structure on the technology set it is supposed to be closed, bounded and convex (Färe and Primont (1995)). Furthermore, it is supposed that standard axioms such as strong disposability of the inputs and the good outputs are satisfied. Two additional axioms are required in the context of an environmental efficiency analysis to incorporate the special role of undesirable outputs in a consistent way. The first is null-jointness, meaning that it is not possible to produce positive quantities of the good outputs without generating emissions (i.e. if $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}) \in \mathcal{T}$ and $\boldsymbol{u} = \boldsymbol{0}$ then $\boldsymbol{y} = \boldsymbol{0}$). The second is weak disposability stating that proportional reductions of emissions are always feasible as long as the good outputs are reduced by the same proportion (i.e. if $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}) \in \mathcal{T}$ then $(\boldsymbol{x}, \alpha\boldsymbol{y}, \alpha\boldsymbol{u}) \in \mathcal{T}$ for $\alpha \in [0, 1]$). For more detailed discussions of these axioms see Färe and Grosskopf (2004), Färe et al. (2005) and Zhou et al. (2008a).[24]

3.3.2 Directional Distance Functions

The directional distance function (DDF) is defined on the technology set $\mathcal{T}$ as proposed by Chambers et al. (1996) and extended to the incorporation of undesirable outputs by Chung et al. (1997). It is a generalization of the Shephard (1970) distance function to the case of non-proportional changes of the inputs and outputs and can be formally stated as

$$DDF(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}; \boldsymbol{g}_x, \boldsymbol{g}_y, \boldsymbol{g}_u) = \sup\{\delta \geq 0 : (\boldsymbol{x} - \delta\boldsymbol{g}_x, \boldsymbol{y} + \delta\boldsymbol{g}_y, \boldsymbol{u} - \delta\boldsymbol{g}_u) \in \mathcal{T}\}. \tag{3.2}$$

Herein, the inefficiency measure δ expresses the distance of a particular input-output combination $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u})$ towards the boundary of the technology set along a particular direction $\boldsymbol{g}_x \geq \boldsymbol{0}$, $\boldsymbol{g}_y \geq \boldsymbol{0}$, $\boldsymbol{g}_u \geq \boldsymbol{0}$. This measure is equal to zero if the input-output combination is a point on the boundary (is on the frontier function) and it is larger than zero if the input-output combination is below the boundary (is below the frontier function).

In the following we mostly impose the restriction $\boldsymbol{g}_x = \boldsymbol{0}$ and $\boldsymbol{g}_y = \boldsymbol{0}$, meaning that the inefficiency is measured exclusively as the extent of possible reduction of the bad outputs. In our application the entities under investigation are sectors in different countries. On such a high level of aggregation it is appropriate to assume that no reduction of input usage is intended. Since we are mainly interested in measuring the maximum potential

[22] In the subsequent discussion of the results we will frequently simply refer to the outputs when we mean the good outputs and to the emissions when we mean the bad outputs.

[23] Here we also include conventional inputs as labor and capital. Related papers such as Picazo-Tadeo et al. (2012) measure eco-efficiency scores by directional distance functions without using inputs.

[24] An alternative to this approach is the so-called by-production approach proposed by Murty et al. (2012) which relies on the availability of abatement options (and requires appropriate data). This approach models the technology set as the intersection of two parts to be estimated separately. One part is related to the production of the good outputs and the other part is related to the production of the bad outputs. This setting avoids the assumptions of weak disposability and null-jointness. Further discussion and critique is provided by Dakpo et al. (2016).

emission reductions, we also exclude the possibility of output enhancement for most of the analysis. In one variant we only impose $\boldsymbol{g}_x = \mathbf{0}$ so that the output enhancement would also be possible.

The data required for the computation of the DDF pertain to n countries in a particular sector. The analysis is performed for each sector separately, so that an additional index to distinguish sectors is not necessary. The data for the m inputs are contained in the $m \times n$ matrix $\boldsymbol{X}$ with the ith column $\boldsymbol{x}_i$ comprising the input quantities of country i $(i = 1, ..., n)$. Likewise, the data for the s good outputs are contained in the $s \times n$ matrix $\boldsymbol{Y}$ and the data for the r bad outputs are contained in the $r \times n$ matrix $\boldsymbol{U}$, with the ith columns $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ comprising the observations pertaining to country i for the good and bad outputs, respectively.

In (3.2) the direction vectors $\boldsymbol{g}_y$ and $\boldsymbol{g}_u$ are not specified. A frequent choice in applications is to make the directions proportional to the variables $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ which serves to let the inefficiency measure be invariant to units of measurement (see e.g. Chung et al. (1997) and Färe et al. (2007)). Since this is restrictive it would be beneficial to compute the directions endogenously. Hampf and Krüger (2015) propose one possibility to endogenize the direction in an environmental efficiency setting and Färe et al. (2013) provide a related proposal to compute endogenous directions in the case of a slacks-based inefficiency measure. As pointed out by Chen and Delmas (2012), these proposals have the additional advantage of avoiding the problem of dominated (weakly-efficient) reference points on the frontier function.

We follow Hampf and Krüger (2015) and propose the following optimization problem to endogenize the computation of the direction vector

$$\begin{array}{lrcrcr} \max\limits_{\delta,\boldsymbol{\alpha}_y,\boldsymbol{\alpha}_u,\boldsymbol{\lambda}} & & & \delta & & \\ \text{s.t.} & \boldsymbol{x}_i & & & \geq & \boldsymbol{X}\boldsymbol{\lambda} \\ & \boldsymbol{y}_i & + & \delta\boldsymbol{\alpha}_y \odot \boldsymbol{y}_i & \leq & \boldsymbol{Y}\boldsymbol{\lambda} \\ & \boldsymbol{u}_i & - & \delta\boldsymbol{\alpha}_u \odot \boldsymbol{u}_i & = & \boldsymbol{U}\boldsymbol{\lambda} \\ & & & \mathbf{1}'\boldsymbol{\alpha}_y + \mathbf{1}'\boldsymbol{\alpha}_u & = & 1 \\ & & & \boldsymbol{\lambda}, \boldsymbol{\alpha}_y, \boldsymbol{\alpha}_u & \geq & \mathbf{0} \end{array} \tag{3.3}$$

where '$\odot$' denotes the direct (Hadamard) product. Herein, $\boldsymbol{\lambda}$ is a n-vector containing the weight factors to determine the reference point on the frontier function. The direction weights $\boldsymbol{\alpha}_y$ and $\boldsymbol{\alpha}_u$ are computed jointly with δ and $\boldsymbol{\lambda}$ with the objective of maximizing the distance towards the frontier function. The identification of δ is permitted by the additional constraint $\mathbf{1}'\boldsymbol{\alpha}_y + \mathbf{1}'\boldsymbol{\alpha}_u = 1$. In this specification the direction vectors are proportional to $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ which lets the inefficiency measure be invariant to the units of measurement.

The optimization problem (3.3) is nonlinear and therefore difficult so solve. This is caused by δ and $\boldsymbol{\alpha}_y$ or $\boldsymbol{\alpha}_u$ arising multiplicatively. By defining $\boldsymbol{\gamma}_y = \delta\boldsymbol{\alpha}_y$ and $\boldsymbol{\gamma}_u = \delta\boldsymbol{\alpha}_u$ the problem can be transformed to a well-behaved linear programming problem

$$\begin{array}{lrcrcr} \max\limits_{\boldsymbol{\gamma}_y,\boldsymbol{\gamma}_u,\boldsymbol{\lambda}} & \mathbf{1}'\boldsymbol{\gamma}_y & + & \mathbf{1}'\boldsymbol{\gamma}_u & & \\ \text{s.t.} & \boldsymbol{x}_i & & & \geq & \boldsymbol{X}\boldsymbol{\lambda} \\ & \boldsymbol{y}_i & + & \boldsymbol{\gamma}_y \odot \boldsymbol{y}_i & \leq & \boldsymbol{Y}\boldsymbol{\lambda} \\ & \boldsymbol{u}_i & - & \boldsymbol{\gamma}_u \odot \boldsymbol{u}_i & = & \boldsymbol{U}\boldsymbol{\lambda} \\ & & & \boldsymbol{\lambda}, \boldsymbol{\gamma}_y, \boldsymbol{\gamma}_u & \geq & \mathbf{0} \end{array} \tag{3.4}$$

Taking the constraint $\mathbf{1}'\boldsymbol{\alpha}_y + \mathbf{1}'\boldsymbol{\alpha}_u = 1$ from (3.3) into account we easily see that the value of the objective function $\mathbf{1}'\boldsymbol{\gamma}_y + \mathbf{1}'\boldsymbol{\gamma}_u = \delta \cdot (\mathbf{1}'\boldsymbol{\alpha}_y + \mathbf{1}'\boldsymbol{\alpha}_u)$ is equal to δ as before. Program (3.4) can be easily solved by the ordinary simplex algorithm.[25] The solution values for δ, $\boldsymbol{\alpha}_y$ and $\boldsymbol{\alpha}_u$ can be backed out from the solutions for $\boldsymbol{\gamma}_y$ and $\boldsymbol{\gamma}_u$ by $\delta = \mathbf{1}'\boldsymbol{\gamma}_y + \mathbf{1}'\boldsymbol{\gamma}_u$ as well as $\boldsymbol{\alpha}_y = \boldsymbol{\gamma}_y/\delta$ and $\boldsymbol{\alpha}_u = \boldsymbol{\gamma}_u/\delta$. For a particular country i the solution values are denoted δ_i, $\boldsymbol{\alpha}_{yi}$, $\boldsymbol{\alpha}_{ui}$, $\boldsymbol{\gamma}_{yi}$, $\boldsymbol{\gamma}_{ui}$ and $\boldsymbol{\lambda}_i$ $(i = 1, ..., n)$.[26]

With these solution values we can compute the efficient input-output combination on the frontier function with the coordinates $\hat{\boldsymbol{x}}_i = \boldsymbol{X}\boldsymbol{\lambda}_i$, $\hat{\boldsymbol{y}}_i = \boldsymbol{Y}\boldsymbol{\lambda}_i$ and $\hat{\boldsymbol{u}}_i = \boldsymbol{U}\boldsymbol{\lambda}_i$. The potential reductions of the r bad outputs for country i in the sector under consideration can be computed as $\boldsymbol{u}_i - \hat{\boldsymbol{u}}_i = \boldsymbol{\gamma}_{ui} \odot \boldsymbol{u}_i = \delta_i \boldsymbol{\alpha}_{ui} \odot \boldsymbol{u}_i$. We see that the potential emission reductions depend on the magnitude of the inefficiency measure δ_i as well as on the optimized direction vector $\boldsymbol{\alpha}_{ui}$ of country i. The total emission reduction potential of country i is the sum over all emission categories $RP_i = \mathbf{1}'(\boldsymbol{u}_i - \hat{\boldsymbol{u}}_i)$ with $\mathbf{1}$ denoting a conformable vector of ones and the prime denoting transposition. The sum can, of course, only be validly computed if the emission variables are denominated in a common unit of measurement. This is indeed the case in our application where greenhouse gas emission are expressed in CO_2 equivalents. To report the results later on we further aggregate the potential emission reductions across countries and sectors. Potential output enhancement can likewise be computed as $\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i = \boldsymbol{\gamma}_{yi} \odot \boldsymbol{y}_i = \delta_i \boldsymbol{\alpha}_{yi} \odot \boldsymbol{y}_i$ for the case where we do not impose $\boldsymbol{\gamma}_{yi} = \mathbf{0}$ or $\boldsymbol{\alpha}_{yi} = \mathbf{0}$ a priori.

3.3.3 Variable Returns to Scale

All above stated optimization problems compute the inefficiency measures under the assumption of constant returns to scale (CRS). In a cross-country sectoral setting with countries of rather different size and with a rather different sectoral structure CRS seems to be an overly restrictive assumption. So it would be beneficial to get rid of this rather unrealistic assumption and to measure inefficiency under variable returns to scale (VRS). In nonparametric approaches of efficiency measurement VRS is usually induced by adding the constraint $\mathbf{1}'\boldsymbol{\lambda} = 1$ to the optimization problems. In the case of environmental efficiency analysis this would violate the weak disposability property. Zhou et al. (2008a) show how to induce VRS in a way which is consistent with weak disposability. This implementation again leads to a linear programming problem

$$
\begin{array}{lrcll}
\max\limits_{\beta,\boldsymbol{\gamma}_y,\boldsymbol{\gamma}_u,\boldsymbol{\zeta}} & \mathbf{1}'\boldsymbol{\gamma}_u & + & \mathbf{1}'\boldsymbol{\gamma}_u & \\
\text{s.t.} & \beta\boldsymbol{x}_i & & & \geq \quad \boldsymbol{X}\boldsymbol{\zeta} \\
& \boldsymbol{y}_i & + & \boldsymbol{\gamma}_y \odot \boldsymbol{y}_i & \leq \quad \boldsymbol{Y}\boldsymbol{\zeta} \\
& \boldsymbol{u}_i & - & \boldsymbol{\gamma}_u \odot \boldsymbol{u}_i & = \quad \boldsymbol{U}\boldsymbol{\zeta} \\
& & & \mathbf{1}'\boldsymbol{\zeta} & = \quad \beta \\
& 1 \geq \beta \geq 0 & , & \boldsymbol{\zeta}, \boldsymbol{\gamma}_y, \boldsymbol{\gamma}_u & \geq \quad \mathbf{0}
\end{array}
\tag{3.5}
$$

with an additional parameter β which is bounded in $[0, 1]$. Details can be found in Zhou et al. (2008a). As before, we obtain the solution values for $\boldsymbol{\gamma}_u$ which allow to back out

[25] For the computation of the solutions in this paper the R-package "lpSolve" is used.

[26] In the case of the efficient countries (with $\delta = 0$) the solution for $\boldsymbol{\alpha}_y$ and $\boldsymbol{\alpha}_u$ is indeterminate. Clearly, there exists no direction towards the frontier function if an observation already stays on the frontier function.

$\delta = \mathbf{1}'\boldsymbol{\gamma}_y + \mathbf{1}'\boldsymbol{\gamma}_u$, $\boldsymbol{\alpha}_y = \boldsymbol{\gamma}_y/\delta$ and $\boldsymbol{\alpha}_u = \boldsymbol{\gamma}_u/\delta$ and to compute the emission reduction potentials. This problem can again be easily solved by the simplex algorithm. Here also, the solution values are denoted δ_i, $\boldsymbol{\alpha}_{yi}$, $\boldsymbol{\alpha}_{ui}$, $\boldsymbol{\gamma}_{yi}$, $\boldsymbol{\gamma}_{ui}$ and $\boldsymbol{\lambda}_i$ for a particular country i $(i = 1, ..., n)$. We stick to the VRS assumption throughout this paper.

3.3.4 Bootstrapping

The inefficiency measures and the derived reduction potentials are estimates from a data sample which are subject to measurement error and therefore stochastic in nature. Frontier function estimation is associated with a further peculiarity. Specifically, the empirical implementation of the linear programming problems (3.4) or (3.5) is based on the observed input-output combinations in the data. This lets the empirically estimated frontier function provide a closer envelopment of the data than the true (unobserved) frontier function. As a consequence, the empirically determined technology set $\hat{\mathcal{T}}_{\text{DDF}}$ underlying the empirical analysis is a subset of the true technology set $\mathcal{T}$, i.e. $\hat{\mathcal{T}}_{\text{DDF}} \subseteq \mathcal{T}$. This leads to downward-biased estimates of the inefficiency measures and the emission reduction potentials. This bias can be substantial and bootstrapping provides a practical way to achieve a correction (see Simar and Wilson (2008, 2011)).

We resort to a bootstrapping approach to compute bias-corrected estimates of the reduction potentials and to establish confidence intervals for these measures. The specific approach pursued here is analogous to the procedure proposed by Simar and Wilson (1998) adapted to the setting of directional distance functions. Compared to the double-bootstrap algorithm of Simar et al. (2012) the chosen approach is more transparent and easier to communicate. The approach of Simar et al. (2012) uses a complicated orthogonal transformation of the data and two smoothing loops which requires the selection of two critical bandwidth parameters instead of one. This bandwidth choice is particularly problematic in small-sample situations. Moreover, the algorithm seems not to be adapted to the inclusion of bad outputs since the direction vector pertaining to the outputs is restricted to be non-negative.

The smoothed bootstrap algorithm adapted from Simar and Wilson (1998) to the DDF setting starts with some preparatory steps. First, the DDF and the optimal directions are computed from the original data by solving (3.5) to obtain $\hat{\delta}_i$ as well as the optimal directions $\boldsymbol{\alpha}_{yi}$ and $\boldsymbol{\alpha}_{ui}$ for all $i = 1, ..., n$. The directions are computed once and kept fixed during the whole procedure. Furthermore, the bandwidth parameter h for the smoothing is chosen as described in Simar and Wilson (2011) where also some R code is provided.

The main part of the bootstrapping algorithm cycles B times through the following steps:

- A bootstrap resample is obtained by first drawing with replacement from $D = \{\hat{\delta}_1, ..., \hat{\delta}_n, -\hat{\delta}_1, ..., -\hat{\delta}_n\}$ which implements a boundary reflection about zero. The result of this step is denoted $\tilde{\delta}_i$ $(i = 1, ..., n)$.
- The smoothing step is performed by adding $h \cdot \varepsilon_i$ to each draw, where the ε_i are independent standard normal draws, thus obtaining $\tilde{\delta}_i + h \cdot \varepsilon_i$ and finally returning $\delta_i^* = \left| \bar{\delta} + (\tilde{\delta}_i + h \cdot \varepsilon_i - \bar{\delta})/\sqrt{1 + h^2/\tilde{\sigma}_\delta^2} \right|$ for all $i = 1, ..., n$ where $\bar{\delta}$ and $\tilde{\sigma}_\delta^2$ denote the sample mean and variance of $\tilde{\delta}_i$ $(i = 1, ..., n)$, respectively.

- These resampled inefficiencies are used to construct the bootstrap resample of the reference points by setting $\boldsymbol{x}_i^* = \boldsymbol{x}_i$, $\boldsymbol{y}_i^* = \boldsymbol{y}_i + (\hat{\delta}_i - \delta_i^*)\boldsymbol{\alpha}_{yi} \odot \boldsymbol{y}_i$, $\boldsymbol{u}_i^* = \boldsymbol{u}_i - (\hat{\delta}_i - \delta_i^*)\boldsymbol{\alpha}_{ui} \odot \boldsymbol{u}_i$ for all $i = 1, ..., n$. By that operation the observation $(\boldsymbol{y}_i, \boldsymbol{u}_i)$ is first projected on the frontier (by $+\hat{\delta}_i$) and then randomly away from the frontier (by $-\delta_i^*$) along the fixed direction ($\boldsymbol{\alpha}_{yi} \odot \boldsymbol{y}_i$ and $-\boldsymbol{\alpha}_{ui} \odot \boldsymbol{u}_i$). The resulting bootstrap resample consists of $\boldsymbol{X}^* = (\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*)$, $\boldsymbol{Y}^* = (\boldsymbol{y}_1^*, ..., \boldsymbol{y}_n^*)$ and $\boldsymbol{U}^* = (\boldsymbol{u}_1^*, ..., \boldsymbol{u}_n^*)$.

- The efficiency measures are computed by solving (keeping the directions fixed)

$$\begin{array}{lrcll} \max\limits_{\beta,\delta,\boldsymbol{\zeta}} & & & \delta & \\ \text{s.t.} & \beta\boldsymbol{x}_i & & & \geq \quad \boldsymbol{X}^*\boldsymbol{\zeta} \\ & \boldsymbol{y}_i & + & \delta\boldsymbol{\alpha}_{yi} \odot \boldsymbol{y}_i & \leq \quad \boldsymbol{Y}^*\boldsymbol{\zeta} \\ & \boldsymbol{u}_i & - & \delta\boldsymbol{\alpha}_{ui} \odot \boldsymbol{u}_i & = \quad \boldsymbol{U}^*\boldsymbol{\zeta} \\ & & & \mathbf{1}'\boldsymbol{\zeta} & = \quad \beta \\ & \beta \geq 0 & , & \boldsymbol{\zeta} \geq \mathbf{0} & \end{array} \tag{3.6}$$

  for each $i = 1, ..., n$, where $\boldsymbol{x}_i$, $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ constitute the original observation for country i and $\boldsymbol{X}^*$, $\boldsymbol{Y}^*$ and $\boldsymbol{U}^*$ are taken from the preceding step. The results are the bootstrap inefficiency measures $\hat{\delta}_i^*$ for all $i = 1, ..., n$. From the bootstrap inefficiency measures the emission reduction potentials $\Delta\hat{\boldsymbol{u}}_i^* = \hat{\delta}_i^*\boldsymbol{\alpha}_{ui} \odot \boldsymbol{u}_i$ or potential output enhancement $\Delta\hat{\boldsymbol{y}}_i^* = \hat{\delta}_i^*\boldsymbol{\alpha}_{yi} \odot \boldsymbol{y}_i$ are obtained for all $i = 1, ..., n$.[27]

Cycling through the preceding steps B times we obtain the bootstrap resamples ($\hat{\delta}_{i,b}^*$, $\Delta\hat{\boldsymbol{y}}_{i,b}^*$, $\Delta\hat{\boldsymbol{u}}_{i,b}^*$) with $b = 1, ..., B$ for each country $i = 1, ..., n$ in a given sector.

Based on the bootstrap resamples the bias correction and percentile confidence intervals can be obtained. Letting z_j be the generic notation of either interesting variable (e.g. aggregates of reduction potentials over sectors or countries), we denote the estimate from the original data by $\hat{z}_j$ and the bootstrap resamples by $\hat{z}_{j,b}^*$ for each $b = 1, ..., B$.

The bias correction is performed by computing $\hat{z}_{\text{bc},j} = \hat{z}_j - \widehat{\text{bias}}_j$ with $\widehat{\text{bias}}_j = B^{-1}\sum_{b=1}^B \hat{z}_{j,b}^* - \hat{z}_j$. This measure is only computed for those cases j (countries or sectors) where $|\widehat{\text{bias}}_j|/\hat{\sigma}_j > 1/\sqrt{3}$ with $\hat{\sigma}_j^2 = (B-1)^{-1}\sum_{b=1}^B (\hat{z}_{j,b}^* - \bar{z}_{j,b}^*)^2$ and $\bar{z}_{j,b}^* = B^{-1}\sum_{b=1}^B \hat{z}_{j,b}^*$. The rationale for this rule is that the absolute bias has to be sufficiently large compared to the standard deviation in order to achieve a reduction in the mean squared error from the bias correction (see Simar and Wilson, 2008, pp. 449f.).

Percentile confidence intervals $[\hat{z}_{\text{cl},j}, \hat{z}_{\text{cu},j}]$ are established using the $\alpha/2$ and $1-\alpha/2$ percentiles of $\{\hat{z}_{j,1}^*, ..., \hat{z}_{j,B}^*\}$, denoted $\hat{z}_{\text{cl},j}$ and $\hat{z}_{\text{cu},j}$, respectively, for some confidence level $1-\alpha$. For the usual value of $\alpha = 0.05$ we thus have $\Pr(\hat{z}_{\text{cl},j} \leq z_j \leq \hat{z}_{\text{cu},j}) = 0.95$.

One problem that occasionally arises during the bootstrap resamples is that we obtain reduction potentials which are larger than the actual emission quantities. To deal with this problem we prune out those cases in the spirit of an accept-reject procedure. The bias correction and the confidence intervals are established from this truncated distribution. Since we base the confidence intervals on a large number of bootstrap replications (actually $B = 20000$)[28] there always remains a sufficient number of replications for obtaining reliable

[27] As a computational detail an offset is added to δ in (3.6) and subtracted after the solution is obtained. This allows for negative values for δ arising during the bootstrap replications.

[28] We also explored an even higher number of 50000 bootstrap replications for selected variants and reached essentially the same findings.

estimates of the confidence bounds. The bias correction, which is more concerned with the center of the distribution instead of the tails, is even less affected by the pruning operation anyway.

3.4 Results and Discussion

The exposition of the results in this section is structured along four specification variants: (a) with a single emission variable (total GHG emissions in CO_2e), (b) with two emission variables (CO_2 and other (non-CO_2) GHG emissions) and (c) with three emission variables (CO_2, CH_4, N_2O measured in CO_2e). Whereas we fix $\boldsymbol{\alpha}_y = \mathbf{0}$ in these variants, we pursue an additional variant (d) where we also allow for enhancement of the (good) output in addition. We consider the two subperiods t_1 = 2008-2012 and t_2 = 2012-2016 where all inputs and outputs are computed as medians over the indicated years to achieve greater robustness of the results.[29] Most of the discussion focuses on the second subperiod, since there is not much change in the pattern of results across sectors and countries between both subperiods and the results for the more recent subperiod are of greater relevance for the current discussion about climate change. All variants include the conventional inputs capital K and labor L as well as the single economic output Y as defined in section 3.2. Finally, in subsection 3.4.4 we conclude this section with an extensive discussion of policy implications.

3.4.1 Total Greenhouse Gas Emissions

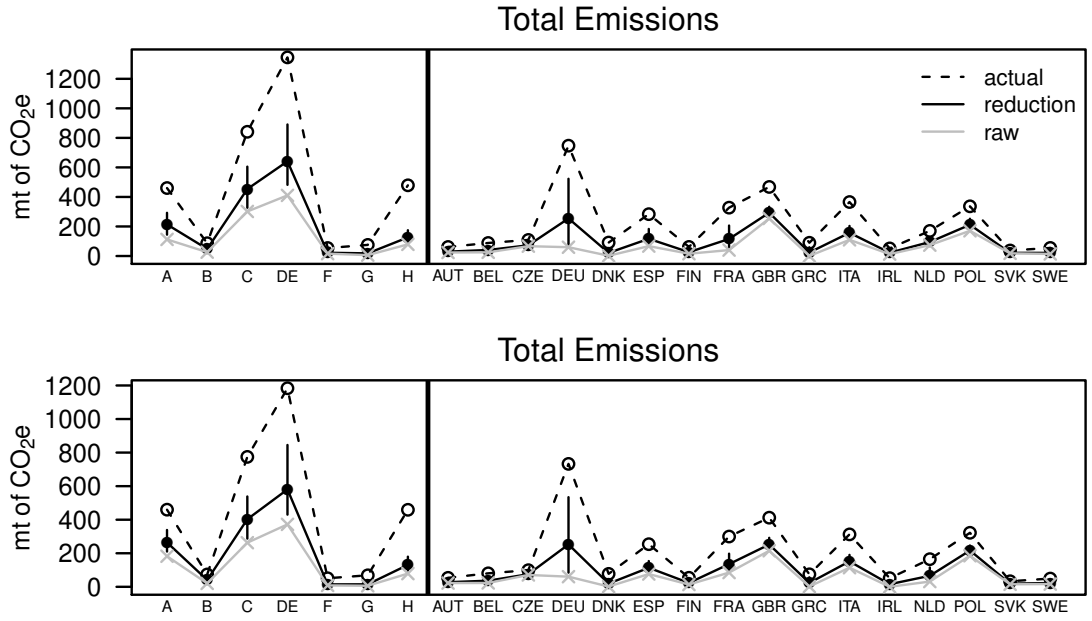
We start with total GHG emissions as the single emission variable. Figure 3.2 depicts the results for the first subperiod in the upper panel and for the second subperiod in the lower panel. In each panel we depict the actual emissions (open circles, connected by a dashed line) and the bias-corrected emission reduction potentials (bullet points, connected by a solid line) aggregated for the 7 sectors on the left side and the 16 countries on the right side. The vertical lines extending above and below the bullet points indicate the 95% bootstrap confidence intervals as explained above. The scale on the ordinate reveals that all values are expressed in million tons (mt) of CO_2e.

The corresponding numerical results are reported in Table A2 in the appendix for reference. There we see that the sum of actual emissions over all sectors and countries amounts to 3341 mt of GHG (in CO_2e) in the first subperiod which decrease to 3070 mt in the second subperiod. Total bias-corrected reduction potentials are 1522 mt (with 95% confidence interval $[1236, 1886]$) in the first subperiod and 1448 mt (with 95% confidence interval $[1177, 1801]$) in the second subperiod. These amount to 46% and 47% of the actual emissions in the two subperiods, respectively.

Figure 3.2 shows that the bias-corrected reduction potentials are quite sizable in some sectors and in a number of countries. Notice that the countries are of rather different size economically and this is also reflected in the differences of the actual GHG emissions. The bias correction lets the estimates of the reduction potentials appear much larger compared to their "raw" equivalents discussed at length in the companion paper of

[29] Recall that the median is a more robust measure of location compared to the mean.

Figure 3.2: Potential emission reduction of total GHG for the period 2008-2012 (upper panel) and the period 2012-2016 (lower panel), variant (a)



Krüger and Tarach (2020).[30] Most of the confidence intervals are quite narrow pointing to rather precise estimates of the potential emission reductions. There are some exceptions, however, where the confidence intervals are wider. These exceptions pertain to sectors or countries with larger actual emissions and larger reduction potentials (e.g. sectors C and DE or Germany).

We first turn to the lower panel with the results for the second subperiod. Regarding the sectors the estimated reduction potentials are particularly large in sector DE (mostly energy) and C (manufacturing) where they amount to 49% and 52% of the actual emissions, respectively.[31] This is followed by sectors A (agriculture) and H (transport) where the reduction potentials are smaller in absolute terms, amounting to 57% and 29% of the actual emissions, respectively. For the remaining sectors the reduction potentials are small to negligible.

Germany (DEU) is the country with the largest actual emissions, but its reduction potential is of the same size as that of the United Kingdom (which is considered as an EU member country during the sample period until 2016) which has the second largest actual emissions. Other countries with sizable reduction potentials are Poland, Italy, France and Spain (with decreasing reduction potentials in this order). The reduction potentials amount to 34% of the actual emissions in the case of Germany, 61% for the United King-

[30]The gray line depicts the "raw" reduction potentials (without bias correction) which are throughout smaller than their bias-corrected counterparts and track them quite closely, except in some cases where the actual emissions are particularly large.

[31]In sector DE emissions are to a large extent determined by the share of fossil power-generation. France and Sweden are very efficient and have shares of nuclear and water power of about 80% (see NEA-OECD (2018), p. 19 and Byman (2016), p. 8, respectively). Thus, the reduction potentials of the other countries implicitly require similar shares of non-fossil power-generation which may be realized by either nuclear power, water power or other renewable forms (biomass, solar, wind).

dom, 67% for Poland, 48% for Italy, 45% for France and Spain. Reduction potentials of the other 10 countries in our sample are absolutely smaller and thus are not as easily visible. Nonetheless, for both groups of countries reduction potentials are very much in proportion to their actual GHG emissions. During 2012-16, these 10 smaller countries account for almost one quarter of total GHG emissions in our sample. As can be verified from Table A2, summing the reduction potentials of these countries results in almost one quarter of the total reduction potential as well. Hence, although we primarily focus our discussion here on the results for the largest countries, we want to stress that the joint contribution of the smaller countries to emission reduction is as important as that of the largest countries.

Comparing these results with those for the first subperiod in the upper panel of the figure we find the distributions across sectors and countries to be rather similar in both subperiods. The major difference is the overall magnitude of the actual emissions and the estimated reduction potentials which are both larger in the first compared to the second subperiod. There are exceptions from this rule in the some sectors and countries as can be seen from a closer inspection of Table A2 in the appendix but these are of minor quantitative importance. Combined with the greater relevance of the more recent subperiod in the current debate about climate change, this similarity justifies the focus on the subperiod (2012-2016) in the subsequent discussion of the other variants

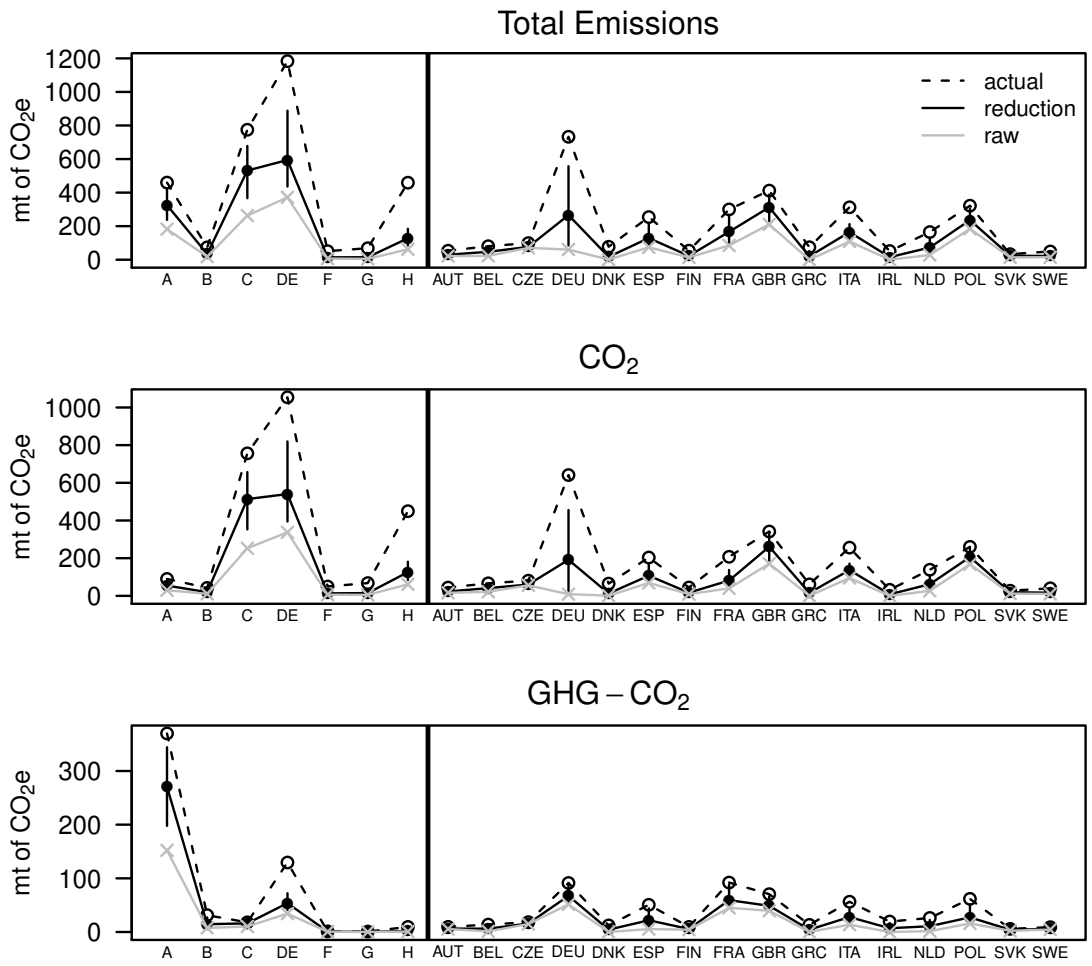
3.4.2 CO_2 and other GHG (CH_4 and N_2O) Emissions

In this subsection, we proceed by splitting total GHG emissions into CO_2 and other GHG emissions (which is the sum of CH_4 and N_2O, or equivalently total GHG minus CO_2, expressed in CO_2e). Bias-corrected estimates of reduction potentials and confidence intervals are calculated for each of the two emission variables as well as for their sum (total GHG emissions). Figure 3.3 shows actual emissions and reduction potentials for the sectors and countries in the sample, and Table A3 in the appendix permits a closer look at the exact quantities. Aggregated over all sectors and countries we find total GHG reduction potentials of 1642 mt (with 95% confidence interval [1321, 2049]), amounting to 53% of the actual emissions.

The lowest panel of the figure reveals that sector A is by far the largest emitter of other GHG, and these emissions also account for the majority of emissions in this sector (about 80%). In addition to sector A, emissions of other GHG are sizable in absolute terms in sector DE, although they amount only to about 10% of the sector's total GHG emissions. In contrast, emissions of other GHG in sector B are small in absolute size, but nonetheless make up 42% of total GHG emissions here. Figure 3.3 shows that the sectors C and H emit mostly CO_2. Together with sector DE they account for the majority of CO_2 emissions in our sample. One result of this subsection, which is evident from the middle panel of Figure 3.3, is that sector C has almost the same CO_2 reduction potential as sector DE, although it emits clearly less CO_2, which points to the relevance of the manufacturing sector for saving CO_2 emissions. Another finding is that there is overall higher inefficiency regarding other GHG rather than CO_2 in relative terms. In particular, the share of reduction potentials to actual emissions for other GHG is with 64% (largely concentrated in sector A) considerably above the corresponding value for CO_2 (51%).

Looking at the countries we observe that reduction potentials for other GHG are particularly sizable for the two largest emitters of other GHG, namely France and Germany,

Figure 3.3: Potential emission reduction of CO_2 and other GHG for the period 2012-2016, variant (b)



where they make up 42% and 26% of the countries' total GHG reduction potentials, respectively. This indicates that it is important also to include the agricultural sector with its CH_4 and N_2O emissions in the respective emission reduction plans in France and Germany. For some other large countries, the potentially feasible reductions for other GHG are comparably smaller than for CO_2 (at or below 17% of total GHG reduction potentials in the cases of the United Kingdom, Poland, Italy and Spain). Confidence intervals are generally quite narrow for reductions of other GHG for most sectors and countries.

Next, we further split the other GHG emissions into CH_4 and N_2O emissions. Figure 3.4 shows the results of this split (see also Table A4). The differences to the results above can be explained by the direction choice where now one further possibility is available. During the second subperiod (2012-2016), our sample countries and sectors emitted a total of 3070 mt CO_2e, out of which 6% were N_2O emissions, 12% CH_4 emissions, and 82% CO_2 emissions. As above, the shares of reduction potentials in actual emissions are higher for N_2O (81%) and CH_4 (57%) than for CO_2 (45%). For total GHG this translates into a share of reduction potentials in actual emissions of 48%, or 1487 mt (with 95% confidence interval $[1214, 1840]$).

Since the use of nitrogen fertilizers is the major source of N_2O emissions, their reduction potentials are clearly largest in sector A. Here, we find that the lower bound of the confidence interval indicates that reduction potentials below 65% of actual emissions are unlikely at the 95% confidence level. Our results imply that there is the potential for the sample countries to continue the N_2O reduction path that has already started in Europe during 1990-2010 (see Tian et al. (2020) for regional trends of anthropogenic N_2O emissions for the period 1980-2016). Sector A also is the major source of CH_4 emissions (mainly from animal livestock) and our results reveal that it has a sizable potential for reducing CH_4. Our bias-corrected estimate is 69% of actual emissions (144 mt of CO_2e) and a reduction potential of less than 50% of actual emissions is unlikely based on the 95% confidence interval. In addition to sector A, Figure 3.4 reveals that sector DE contributes substantially to CH_4 emissions with a rather small reduction potential. Finally, for sector B the CH_4 reduction potentials are an order of magnitude smaller than those of sector A, and for the other sectors they are negligible.

3.4.3 Combined Direction with Output Enhancement

In the preceding subsections we have measured inefficiency exclusively in the direction of reducing emissions, holding the good output and the inputs constant. As noted in section 3.3.2, it is not sensible to measure inefficiency in the direction of reducing the inputs at this high level of aggregation. However, instead of pure emission reductions a combination of emission reductions and output enhancement (i.e. economic growth) is a realistic objective of policy makers. We account for this possibility by allowing more flexible directions $\boldsymbol{\alpha}_y \geq \mathbf{0}$ in the linear program of equation (3.3).[32] By this, we allow emission reductions to be traded off against output enhancement, so that inefficiency measures reflect a combination of emission reduction and output enhancement.

These estimates are reported (again for the second subperiod) in Figure 3.5 and Table A5. For the emission variable we use total GHG, so that reduction potentials can be directly compared with those from subsection 3.4.1 for the period 2012-16. In total, allowing for

[32]The boundary solution of pure emission reductions is still possible but needs no longer be optimal.

Figure 3.4: Potential emission reduction of CO_2, CH_4 and N_2O for the period 2012-2016, variant (c)

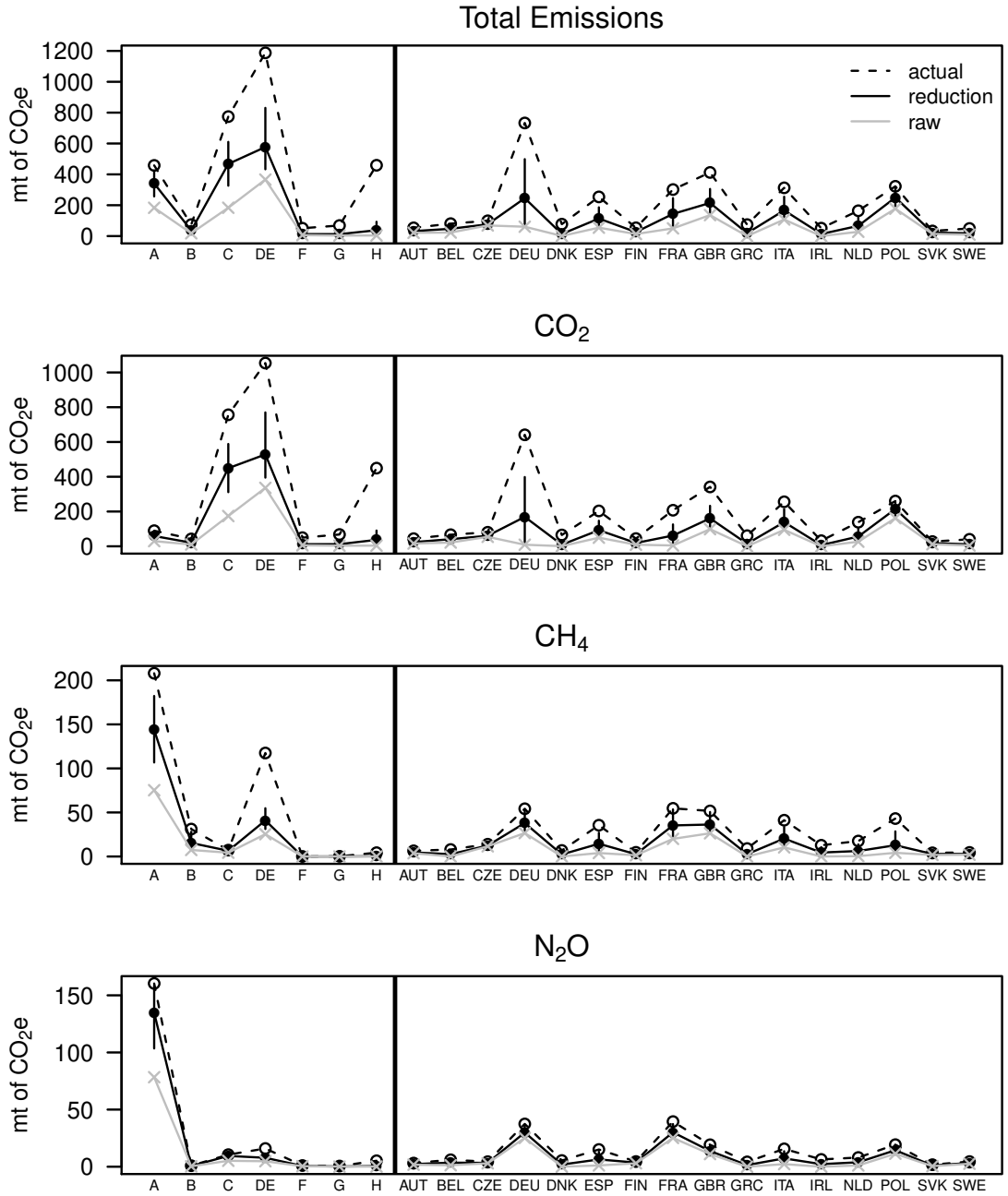
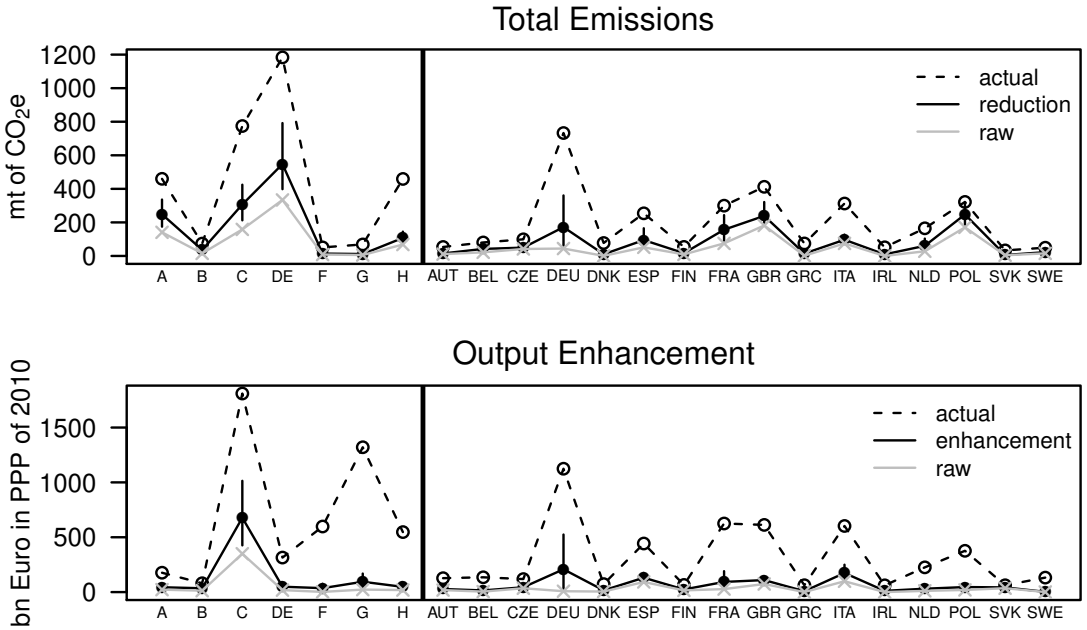


Figure 3.5: Potential emission reduction of total GHG and output enhancement for the period 2012-2016, variant (d)



output enhancement causes reduction potentials to decline from 1448 mt to 1272 mt (with 95% confidence intervals changing from [1177, 1801] to [1028, 1595]), or equivalently from 47% to 41% of the actual emissions. As this decline is rather small, the result indicates that most inefficiency is due to generating too much emissions instead of producing too less of the economic output. When comparing the upper panel of Figure 3.5 with the lower panel of Figure 3.2, we notice only slight differences regarding the distribution across sectors. It is visible that reduction potentials for sector C (manufacturing) decline the most (by about 100 mt) from 52% of actual emissions to 40%. For the other sectors the differences are much smaller. This is in line with our estimates for output enhancement potentials, which are low in all sectors except C (see Figure 3.5, lower panel).[33] Hence, we find that a great deal of inefficiency in the direction of output enhancement, being large enough to substantially lower the scope for emission reductions, is present only in sector C.

For most countries we find reduction potentials of somewhat smaller size than in subsection 3.4.1. Figure 3.5 (upper panel) reveals a distribution across countries which is very similar to that of Figure 3.2 (lower panel). There are three exceptions (Belgium, France and Poland) where we obtain even higher reduction potentials than in subsection 3.4.1. This can be attributed to the bias correction applied to the "raw" estimates.

[33]Even for sector G (mainly wholesale and retail trade), which produces substantial value added, output enhancement potential is quite low, indicating that this sector is overall quite efficient in transforming inputs (labor and capital) to output (value added).

3.4.4 Policy Recommendations

Altogether, the preceding discussion shows that there is a sizable extent of inefficiency in the sectors which could be transformed into emission reductions. Therefore, the question arises how can policy support the realization of the potentials in practice. Policies should aim at inducing firms to improve their productive efficiency (which then improves the efficiency of the sectors we measure here) to move towards the frontier function combined with specific measures to channel this movement in the right direction, i.e. mainly in the direction of emission reduction and less in the direction of output enhancement. In the long-run, effort to generate technological progress to shift the frontier function towards less emission intensive modes of production would also work in the right direction, although this aspect is beyond the scope of the measurement exercise we conduct here.

As a theoretical guidepost for structuring the discussion of policy measures we use the macroeconomic model of Stern and Kander (2012). From the first-order condition in equation (15) of that paper one can derive an equation for energy demand as

$$E = \gamma_E^{1/\sigma(1-\phi)} \cdot Y \cdot \left(\frac{p_E}{p}\right)^{-1/(1-\phi)} \cdot A_E^{\phi/(1-\phi)} \tag{3.7}$$

with E denoting the energy input in the underlying CES production function, Y gross output, and p_E/p the price of energy relative to the price of output. A_E is an augmentation factor of the energy input in the production function. This factor increases if a firm or sector in a country improves its energy efficiency. It can be viewed as also comprising the aspect of energy quality as discussed in Stern and Kander (2012) and Stern (2010). Increasing energy quality acts like a further factor (named Q in Stern and Kander (2012)). If we consider E as the input of mostly fossil fuels, changing the energy mix towards renewable energy sources operates analogous to an improvement of energy quality associated with less GHG emissions. The parameter ϕ is related to the elasticity of substitution σ between energy and a capital-labor aggregate by $\phi = (\sigma - 1)/\sigma$ and γ_E is a further production function parameter related to the energy input.

Various estimates of the parameters are reported in table 1 of Stern and Kander (2012). They can be roughly summarized as $\gamma_E \approx 0.2$, $\sigma \approx 2/3$ and $\phi \approx -1/2$. Taken at face value we get a calibrated version of the energy demand equation as

$$E = 0.2 \cdot Y \cdot \left(\frac{p_E}{p}\right)^{-2/3} \cdot A_E^{-1/3}. \tag{3.8}$$

This shows that the energy demand and therefore GHG emissions (assuming that a substantial part of energy is generated from fossil sources) are linearly related to the level of gross output. Emissions are decreasing with a higher relative price of energy p_E/p as well as with an improvement of productive efficiency leading to a larger A_E. These variables are linked to various policy measures which are discussed in the following. In addition to the direct effect of p_E/p there is a further, indirect, effect working through A_E, i.e. adopting more energy efficient and therefore less emission intensive technologies in the medium to long run when p_E/p is higher.

To increase the productive efficiency of firms and sectors, a large variety of policy measures can be beneficial. Examples of these measures directed generally towards inefficiency

reduction are fostering competition, reducing some regulations, incentivizing research and development and protecting intellectual property. Yet, the effect of these policy measures has to be channeled towards improvements in energy efficiency (A_E) or energy quality which are associated with emission reductions. There are many diverse policy measures available which could be appropriately combined and coordinated.

This policy mix, of course, has to comprise traditional instruments of environmental policy such as phasing out the most emission-intensive modes of power generation (i.e. coal), abandoning certain modes of travel (e.g. short-distance flights) or setting energy consumption standards for new products (e.g. cars, heating). Incentives or subsidies for behavior modification (i.e. switching to electric cars, more attractive public transport, thermal insulation of buildings) could also direct efficiency improvement in the desired direction.

A well-designed carbon price is a key measure to influence the relative price of fossil-fuel based energy (p_E/p) and thus, by the energy demand relation, also combustion thereof (Sen and Vollebergh (2018), Best et al. (2020)). Important for the design here is a broad sector and country coverage (Aldy et al. (2010), p. 928). One finding of our efficiency analysis, in particular from subsection 3.4.2, is that emission reduction potentials for the manufacturing sector are substantial, i.e. of similar magnitude as for the energy sector. We also find substantial reduction potentials in the transport sector, although these are smaller in relative terms. Currently, European emission trading still covers only power generation, heavy industry and intra-European aviation (Delreux and Ohler (2019), p. 7). Thus, comprehensive EU-wide carbon tax policies could push manufacturing and transport sectors closer to their frontier in the desired direction of reducing CO_2 emissions. An extension of emission trading to these sectors, if designed effectively (i.e. with a steadily decreasing cap of overall emissions), would also work in the desired direction. However, for the manufacturing sector, we also find in subsection 3.4.3 that there is considerable inefficiency in the output direction. Therefore, a uniform carbon price could easily be set too low in order to channel the direction of inefficiency reduction towards CO_2 reduction in those sectors where much efficiency can also be gained by output enhancement instead.

Furthermore, innovations are important for decarbonization. Sufficient carbon pricing works in the direction of inducing innovations for technologies with high potential for GHG emission reductions and therefore increases A_E (Aghion et al. (2016), van den Bergh and Savin (2021)). These include in particular the new renewable energy technologies (i.e. wind and solar), large-scale and cost-efficient energy storage, and electric cars. Equally important may be all types of innovations which are able to substantially improve the energy efficiency of products or production processes, in particular if the energy used there is still fossil-fuel based.[34] To induce innovative activity, public funding of research and development could be specifically targeted at these areas. As knowledge creation is a public good, there is considerable under-investment in these areas to be expected (Knopf et al. (2013), p. 233). However, it seems generally difficult to anticipate which technologies will be most crucial for the energy system of the future. Thus, with respect to the direction of research and development funding, "policy-makers have to strike a delicate balance between supporting promising developments whilst avoiding the temptation to prematurely pick winners" (Knopf et al. (2013), p. 234).

[34] An example are activities such as steel or cement production, where substitution of fossil fuel may be more difficult and where there are sizable regional differences in energy efficiency (Oda et al. (2012)).

These policy measures appear to be well adapted to the sectors C, DE and H with a large amount of CO_2 emissions, while sector A (agriculture) is special in its emission mix caused by the use of fertilizes, animal livestock, as well as its essentiality for human nutrition. When comparing our results for N_2O to other studies, these seem to be quite optimistic. Winiwarter et al. (2018) estimate that if only the lowest-cost N_2O abatement measures were implemented, global emissions thereof could be reduced by 6.2% compared to a baseline scenario, while if also high-cost abatement measures were implemented, global N_2O reduction potentials would amount to 26%. Independent of the exact quantity of N_2O reduction potentials, there are several N_2O abatement measures available which in the agricultural sector generally comprise any measures that improve the efficiency of nitrogen application to crops (Winiwarter et al. (2018), table 1). There is also evidence that organic farming, in particular biodynamic farming, reduces N_2O emissions per yield and is associated with a modest uptake of CH_4 (Skinner et al. (2019), p. 4).

The extent of measured inefficiency is, of course, not confined to Europe. In other countries, emerging economies in particular, there are also huge quantities of potential emission reductions to be expected which could be realized by means of efficiency improvements. Thus, strengthening international cooperation and technology transfer from advanced countries towards emerging countries at favorable conditions could tremendously enhance the overall benefit.

3.5 Discussion and Conclusion

The bottom line of the results of the above analysis with the stochastic nonparametric approach to environmental efficiency analysis is that the bootstrap bias-corrected estimates of potential reductions of GHG emissions by reducing the inefficiency relative to the most efficient countries in each of seven sectors are quantitatively substantial. Along with the bias-correction also confidence bounds are established which show that the reduction potentials are estimated with great precision in some sectors while being wider in other sectors. The comparison with the results of the companion paper Krüger and Tarach (2020) shows that the bias correction leads to substantially increased estimates of the potential emission reductions.

To put the magnitudes of the estimated potential emission reductions into perspective we compare them to the emission reduction targets recently tightened by the European Commission (see EU (2020)). Therein, an emission target of 45% (meaning a reduction by 55%) compared to the GHG emission levels in 1990 until 2030 is postulated. Since the 16 sample countries of our study overall emitted roughly 5000 mt of GHG in 1990 (retrieved from the EEA greenhouse gas data viewer) this implies a necessary emission reduction of about 2750 mt until 2030. As stated by the European Environmental Agency (EEA (2019)), it is highly likely that the target level of a reduction of 20% compared to 1990 is to be achieved or is even slightly outperformed by 2020. Thus, an emission reduction of about 1000 mt is already achieved to date. The remaining 1750 mt until 2030 are not far away from the bias-corrected estimates of the total potential emission reductions aggregated over both countries and sectors (ranging from 1271 mt in variant (d) to 1642 mt in variant (b)) and are well within the confidence intervals. The distribution of the EU reduction targets across sectors also corresponds to the pattern of our estimated reduction potentials. Therefore, becoming more efficient can provide a substantial part of the reductions until 2030, especially in the sectors with large emission volumes.

The distribution of the reduction potentials across countries reflects their (economic) size. Naturally, larger countries tend to have larger manufacturing and transportation sectors which goes along with a need for a larger energy generating sector. These are the sectors with the largest actual emission quantities and also with the largest reduction potentials. Poland seems to be an exception operating with a much higher emission intensity.[35] Agriculture is another sector with large GHG emission which are here more caused by CH_4 and N_2O rather than CO_2. Thus, in addition to conventional environmental policy measures and new technologies which come to mind first, structural change in a direction towards the less emission intensive sectors could contribute to the realization of reduction potentials. Exactly this form of structural change is taking place since several decades in the form of tertiarization (where tertiarization means an increasing share of the less emission intensive service sector at the expense of the primary and secondary sectors; Fourastié (1949)).

Of course, it is not realistic to expect that these emission reductions can be fully achieved within the next decade. The reasons for this assessment are manifold. Policy measures become effective with a time lag. Despite structural change going in the right direction it is a rather sluggish process taking place over longer spans of time. Here, the demand side and the slowly changing consumer preferences play a major role. Economic actors also adapt to changing conditions such as prices and also adverse reactions are to be expected (such as "rebound effects"; Greening et al. (2000)). Structural change is also impeded by the specific roles of the countries in the context of international specialization which prevents that all countries will reach the same sectoral structure (de Araújo et al. (2020)).

On the other hand, the present analysis is purely static (mostly confined to the medians to 2012-2016) and does not take account of the emission reducing effects of technological change. New technologies and in particular less emission intensive forms of energy generation and mobility are crucial for reaching the targets. Since the European countries are only responsible for a small share of global GHG emissions, spillover effects and the transfer of these emission reducing technologies to countries outside Europe is the key factor for new technologies to become effective for large scale emission reduction. This leads to the natural extension of this work towards a dynamic analysis (by projecting the potential emission reductions into the future) or an extension with a global country sample.

[35] This is in broad agreement with the literature on eco-efficiency, where also some countries as Austria, Germany and Sweden are found to be rather efficient while eastern European countries such as Poland appear very inefficient (see Camarero et al. (2014) and Kortelainen (2008)).

4 Bottom-Up Aggregation of Field-Level Oil Production Profiles via a Successive Sampling Discovery Model and a Birth Process: An Application to the Gulf of Mexico and Norway

4.1 Introduction

Since the dawn of the oil and gas industry, the capacity to maintain or grow the supply of these fuels has periodically regained attention and concern. After oil and gas have become predominant energy sources for the world since the post WWII era, the question of their future availability and affordability has undergone recurrent debates.

The issue may be structured by separating it into two questions based on the distinction between stocks and flows: first, "How much recoverable oil [or gas] exists?", and second, "Which path will production take over time?" (Brandt (2010), p. 1). The second issue, the flow, is what is ultimately relevant, since energy flows are what economies and societies depend on (Jakobsson et al. (2012), p. 861). Though obviously the flow depends on the existing stock, the answer to the second question is not fully determined by the answer to the first one, that is, they are not two sides of the same coin. Discussions in the debate often ignore this or are based on the assumption that production rates will be kept constant until "all" of the resource is exhausted, as implicit in calculations based on resource-to-production ratios. Such an oversimplified model likely gives unrealistic predictions (e.g., Bartlett (2000); Bentley (2002)), and "may induce unwarranted optimism" (Bardi and Lavacchi (2009), p. 647).

On the other hand, as argued by Adelman and Lynch (1997), "undue pessimism" may follow from how the first question is formulated, that is from the fixed view on the resource stock. According to Adelman (1997), mineral and fossil fuel production should rather be understood as a "struggle between depletion and increasing knowledge" (p. 13), since the recoverable stock is a variable quantity that depends on technology, geological knowledge, and economic factors.

To disentangle the issue for the case of oil, Brandt (2010) reviews mathematical modeling approaches that have been used to convert given estimates of the stock of oil into future rates of production. Overall, the models are divided into four categories: top-down models, system dynamic models, bottom-up models, and economic models based on optimal depletion theory. Top-down approaches directly start with modeling the production rate at some aggregate (e.g. national, global) level, while omitting the details how this aggregate emerges from underlying physical and economic mechanisms or from the dozens to thousands of individual production units. The most famous top-down approach is known as the Hubbert curve. Essentially, this is a curve fitting model applied to the observed production or discovery path of a resource, where the parametric curve being fit is bell-shaped (usually the Gaussian function or the derivative of the logistic function), and where the total amount of recoverable resources is either exogenously given or estimated from the fitting.

The model has been first proposed by M. King Hubbert (1956; 1962; 1982), primarily based on empirical considerations about the extraction of oil or other non-renewable resources. The implicit assumptions of the Hubbert curve are the following: that maximum

production (the peak) occurs when approximately half of the resource is depleted, that the decline in production mirrors the increase in production, and that production occurs in a single cycle without multiple peaks (Brandt (2007), p. 3075). Other, more flexible curve types for describing or predicting the dynamics of energy systems at a top-down level are reviewed, e.g., by Höök et al. (2011). Apart from Hubbert's forecasts, the most prominent applications of bell-shaped curves to forecast the global production of conventional oil are from the petroleum geologists Colin Campbell and Jean Laherrère, e.g. in Campbell and Laherrère (1998). Bentley (2023) has recently reviewed the methodology from Campbell and Laherrère based on which their conclusions about peak oil supply were reached. Other applications of Hubbert curves to forecast conventional oil production can be found, for example, in Maggio and Cacciola (2009), Nashawi et al. (2010), Wang et al. (2011), and Ebrahimi and Ghasabani (2015).

The use of such curve fitting models at aggregate levels where several have generated forecasts of an imminent peak in global conventional oil production has repeatedly generated controversy, and commentators seem divided into two groups of those who are "concerned" or "unconcerned" about the "future availability and affordability of oil" (Jakobsson et al. (2012), p. 861). Already in the year of publication of Campbell and Laherrère (1998), the opposing viewpoints were critically examined, for example, by Smil (1998). Recently, Bardi (2019) has discussed aspects of the debate on "peak oil" and the cycles it has undergone during the last 20 years in scientific or press articles including a decline in coverage in recent years.

Declining attention to the issue of oil (and gas) depletion may also be related to the plans of many nations to drastically ramp up their share of renewables in primary energy consumption in light of the recent agreements on climate change mitigation such as the Paris Agreement, together with targets of some nations (e.g. European) to phase out fossil fuels completely. Several authors argue, though, that the anthropogenic climate change problem and the fossil energy problem are interwoven and require a holistic solution, or that depletion of fossil fuels could constrain CO_2 emissions as to limit global warming to a manageable extent (Höök and Tang (2013); Nel and Cooper (2009); Kharecha and Hansen (2008)). Furthermore, despite the recent rise of the new renewable energy technologies (i.e., solar, wind, modern biofuels), the author of this paper is of the opinion that the supply of oil and gas remains crucial for many decades. The reasons for this are multifaceted. First, energy demand in many developing countries is expected to rise as these are growing in terms of GDP per capita and aiming to catch up with the developed countries. In 2021, fossil fuels (oil, gas, coal) made up 82.3% of global primary energy consumption, while renewables (excluding hydropower) made up only 6.7% (BP (2022), p. 9). Also, oil and gas power many critical activities such as transportation and the operation of heavy machinery in mining and agriculture, and oil and gas is used as chemical feedstocks for plastics and fertilizers (Day et al. (2018)). Oil accounted for 90% of the energy consumed globally in the transport sector in 2021 (IEA (2022), p. 146). Moreover, gas-fired plants are the most flexible conventional plants for balancing a rising share of intermittently generated renewable power.

Besides, there are several interrelated difficulties associated with phasing out the current work done by fossil fuels and replacing it with the new renewable energy technologies (e.g. wind turbines, solar panels, batteries, hydrogen-based technology). In particular, these include the high intermittency of renewable power generation, the high requirements of potentially scarce metals (e.g. lithium, cobalt, platinum, and Rare Earth elements), es-

pecially for large-scale battery storage as a means to balance the intermittency (Michaux (2021; 2023)), and also a relatively lower power density than fossil fuels (Smil (2015)). Besides, the new renewables might have a lower energy return on energy invested (EROI) compared to fossil fuels (Hall and Day (2009); Hall et al. (2014)), although several EROI analyses available in the literature have been criticized in terms of methodological rigor (Raugei (2023)), and the issues of boundaries and methodological harmonization for comparative EROI analyses are still being discussed (Raugei (2019); Murphy et al. (2022)). In total, though, the issues mentioned make it less clear in which time frame and to what extent the phasing out will be accomplished, and together with the issue of peak oil (or gas), this makes the problem of predicting the future oil and gas availability an important research topic.

In this paper, we formulate a stochastic bottom-up model for the rate of oil production in a region by combining a size-biased sampling model of the discovery process, a pure birth process model for the discovery times, and empirically founded field-level production profiles. The size-biased sampling model is also known as the Barouch-Kaufman model, named after Gordon Kaufman and his co-workers, including Eytan Barouch, who first studied the model in Kaufman et al. (1975) and Barouch and Kaufman (1976a). We estimate the parameters of the size-biased sampling model for three different regions which, upon combining with the other parts of the bottom-up model, allows to evaluate how well the overall model can predict the actual paths of the discovery and production rates.

Regarding the parameter estimation of the size-biased sampling model, we contribute (in sect. 4.3.2) by making two extensions to the usual expectation-maximization (EM) algorithm based estimation procedure and the usual covariance matrix of the estimators (see, e.g., Nair and Wang (1989); Lee (2008)). We first extend both the estimation procedure and the covariance matrix for the case that the number of fields is random with a Poisson prior distribution (based on Lee (2008), pp. 195f.), and secondly, for the case that the support of the field-size distribution is truncated to a compact interval (based on Dempster et al. (1977)).

Among others, Michel (2011) has already formulated a stochastic model for the oil production in a region which is based on the idea of bottom-up aggregation of field-level production profiles. For this Michel (2011) assumes a certain statistical distribution where the launching times of the fields depend on size so that large fields tend to produce first, but he does not incorporate the Barouch-Kaufman model. In contrast, Jakobsson et al. (2012) show theoretical simulation results from a bottom-up model where the discovery of new fields is explicitly modeled via a variant of the Barouch-Kaufman model. Jakobsson et al. (2012) further show that one can endogenize within the model the drilling of unsuccessful exploration wells by including an "empty area", as also formalized in Herbert (1983), so that the model is extended into a model of the discovery yield per exploration effort.

In this paper, we contribute (in sect. 4.3.3) by proposing two different approaches that allow to integrate the size-biased sampling model with a stochastic model for the discovery times of the fields. We formalize this using the stochastic model called a pure birth process. The first approach estimates a distribution function for the future discovery times directly from extrapolating the past trend of "when" the discoveries have occurred with a suitable trend function. Mathematically, the future discovery times can here be described as coming from an inhomogeneous Poisson process. The second approach is based on the

extension from Jakobsson et al. (2012) or Herbert (1983), combined with a suitable trend function for future exploration wells. Mathematically, the future discovery times can here be described as coming from a more general inhomogeneous pure birth process. We also contribute (in sect. 4.3.4) by showing how a result from Pitman and Tran (2015) can be used to derive for both approaches an asymptotic approximation to the mean function of the discovery and the production rate.

Since the approach of combining a model in which larger fields tend to be found first with a model for field-level production is close to the approaches of Michel (2011) and Jakobsson et al. (2012), it is not surprising that several results we obtain in this paper are qualitatively similar (e.g. peak production often happening before 50% of cumulative production). However, the results from Michel (2011) are not built formally on the Barouch-Kaufman model, and the results from Jakobsson et al. (2012) remain theoretical, i.e. the parameters are not adapted to the discovery history of a particular region. We therefore also contribute to the literature on the modeling of oil supply with our empirical application.

This paper proceeds as follows: we start in sect. 4.2 with a brief review of the literature on models of oil production, structured along the categories from Brandt (2010). Sect. 4.3 introduces the mathematical model, where sect. 4.3.1 begins with field-level production profiles adapted from IEA (2008). Sect. 4.3.2 presents formally the central component, the discovery process model based on size-biased sampling, and explains the strategy we use to estimate its parameters. Sect. 4.3.3 explains the two approaches with which we integrate the size-biased sampling model with a model for the discovery times, and sect. 4.3.4 summarizes the mathematical properties of the overall model. In sect. 4.4, we briefly describe the data sources for our applications to the Gulf of Mexico and Norway. Then, sect. 4.5.1 starts with a descriptive analysis, followed by sect. 4.5.2 where we present the estimation results of the size-biased sampling model. In sect. 4.5.3 we present and discuss the Monte Carlo simulation results of the overall model, and finally sect. 4.6 concludes.

4.2 Literature Review of Mathematical Modeling Approaches

Using data from a large number of oil producing regions, Brandt (2007) conducts an empirical analysis of the fit of the Hubbert curve vs. other parametric curves. In total, Brandt (2007) collects production data from 139 regions at widely varying scales (U.S. states or regions, nations, (sub-)continents). Single-peaked symmetric and asymmetric curves are fitted to the historical production data. Six larger regions representing 36% of 2004 production are classified by the author as borderline cases meaning that the production pattern is judged by the author as not well-characterized by a single-peaked curve (Brandt (2007), p. 3082).[36] Comparing the fit of three different symmetric curves (linear, bell-shaped, exponential) for the remaining regions, the bell-shaped Hubbert curve overall performs best. Extending this comparison study to include also asymmetric versions of the three curves, Brandt (2007) finds that the asymmetric exponential curve has the highest number of regions where it performs best. Overall, Brandt (2007) concludes from his study that symmetric models are not fully satisfactory in describing historical production. However, the author considers the implication of this result for prediction efforts to

[36] These are: Central, Southern, and Western Asia; Saudi Arabia, Venezuela, and Former Soviet Union (Brandt (2007), p. 3078).

be less clear, since fitting asymmetric curves becomes impossible when an evident peak does not (yet) appear in the production record.

Rehrl and Friedrich (2006, pp. 2414ff.) give a stylized description about how discoveries proceed which suggests a bell-shaped pattern as a reasonable approximation. Geological information about the prospects in a region is obtained in proportion to either cumulative discoveries or cumulative exploration effort, and technical know-how specific to the region may be gained alongside. As both information is accumulated, the chance of successful discovery increases, which the authors refer to as the "information effect" (see also Uhler (1976); Reynolds (1999)). This constitutes a positive feedback-loop for the rate of discovery. The negative feedback-loop is the "depletion effect" that the declining amount of oil in the ground makes the remaining oil resources harder to find, the discoveries smaller, etc.. If both feedback-loops are understood mathematically as linear proportionalities, the result is the logistic function. The theory of the information and depletion effect is studied in further detail by Reynolds (2001; 2002) and other papers from the author.

Another framework from which bell-shaped production curves can emerge is described by Bardi and Lavacchi (2009), Bardi et al. (2011), and Perissi et al. (2021). The authors propose as a framework for the extraction of a non-renewable or slowly renewable resource a modified version of the Lotka-Volterra (LV) model. The LV model originates as a model of dynamical predator-prey interactions in ecological systems. Bardi and Lavacchi (2009) explain how the two differential equations of the LV model can give a plausible description of the life-cycle of the extracted resource stock ("the prey") jointly with the life-cycle of the capital stock with which the resource is produced ("the predator") when the reproduction rate of the resource stock is set to zero. The model contains three parameters plus the initial stocks of resource and capital. Bardi and Lavacchi (2009) obtain parameter estimates by fitting the model to the observed paths of the resource and capital stocks for a number of data sets: gold mining where the number of miners is the capital stock proxy, whaling in the 19th century where ship tonnage is the capital stock proxy, and oil discoveries in U.S. lower 48 states and in Norway where the number of exploration wells is the capital stock proxy. Overall the fit is quite good, but the authors add the qualification that the number of cases for which the model could be tested is limited, mainly due to difficulties of obtaining data on reasonable capital stock proxies (Bardi and Lavacchi (2009), p. 652). Also, Bardi et al. (2011) and Perissi et al. (2021) explain how the LV framework can be interpreted as a simple dynamical model of the energy return on energy invested (EROI) or the net energy as the key underlying variable.

The cornerstone of the optimal depletion theory is the classical Hotelling (1931) model. This and various modifications have been used in economics to model the extraction of an exhaustible resource such as oil (e.g., Lin (2009); Anderson et al. (2018)). In the classical model, the depletion of a resource by economic agents is formulated as an inter-temporal optimization problem, where the non-renewable nature of the resource is taken into account by a fixed constraint on the amount of extractable resource over all periods. The resource price is usually an endogenous variable determined by equilibrium of the supply function derived from the inter-temporal optimization with an exogenously given demand function. The basic prediction of the model is that the shadow price of the resource grows at the rate of interest, which under the simplest assumptions translates into the resource price growing at the rate of interest (known as the Hotelling rule), and production declining accordingly as dictated by the demand function. The are many extensions of the basic Hotelling model (see Krautkraemer (1998) for a review). However, most extensions do not

consider time-varying constraints on the production rate, and do not take into account the nature of the discovery cycle or geophysical production characteristics. Exceptions can be found, for example, in Pindyck (1978) and more recently in Anderson et al. (2018) and Holland (2008). The "site development model" of Holland (2008, pp. 69-74) is based on optimal depletion theory but is conceptually also a bottom-up model. Therein, the inter-temporal optimization solves in addition to production from existing sites also how large the new sites which are chosen to be developed next should optimally be.

The approach of bottom-up modeling has been recently reviewed by Jakobsson et al. (2014), who judge it as promising for "identifying areas of uncertainty and new research questions" (p. 113). Bottom-up modeling starts with defining a smallest production unit (e.g. an oil field), and then is concerned with modeling (i) the production from the existing units, (ii) the addition of new production units from the reserves, and (iii) the addition of new reserves (Jakobsson et al. (2014)). As the production behavior of individual oil fields shows well-understood regularities which can be approximated by standard production profiles that depend only a few variables (e.g., ch. 10 of IEA (2008); Höök et al. (2009b; 2009a); Höök (2014); Miller and Sorrell (2014)), a model for the production from existing fields can be derived from these insights. At some point the discovery of new reserves by exploration will be subject to the "depletion effect" or "the law of diminishing returns", causing yield per exploration effort to decline. One approach to model this are so-called "creaming curves" which extrapolate cumulative discoveries towards an asymptote (Meisner and Demirmen (1981); Sorrell and Speirs (2010), p. 220).

The approach which we pursue in this paper is to model the "depletion effect" by a discovery process model at the level of individual fields, for which the most well-known approaches are based on Arps and Roberts (1958) and Barouch and Kaufman (1976a). The Arps-Roberts model assumes a discrete field-size distribution, and postulates that one more unit of exploration effort causes new discoveries in a certain size class in proportion to the number of undiscovered fields in that class times the average areal extent of a field in that class. The result of this assumption are exponential decline curves for the number of undiscovered fields in each size class (e.g., Herbert (1982; 1983); Schuenemeyer and Drew (1983)). Kaufman et al. (1975) and Barouch and Kaufman (1976a; 1976b) were the first to postulate, in mathematical form, that discovery of oil reservoirs may be described by sampling without replacement and proportional to size from a finite population, which again is a random sample from an underlying superpopulation distribution (Lee (2008), p. 17). According to the originators the advantages of this model are that, first, it is based on "explicitly stated postulates [that] can be empirically validated using observed data", second, it provides "not only single-number estimates, but an explicit measure of the degree of uncertainty" (Kaufman et al. (1975), p. 13).

Finally, let us turn to the bottom-up models mentioned at the end of sect. 4.1 which are closest to the model of this paper. Michel (2011) formulates a stochastic model where oil production in a region is derived from the combination of field-level production profiles, the field-size distribution, and the dates when the fields are launching oil production. Michel (2011) assumes a truncated Pareto distribution as the field-size distribution, and that the launching times are conditionally (on the field sizes) Gamma distributed where the parameters vary for four different size classes so that larger fields tend to be launched first. A novel and interesting approach pursued by Michel (2011) is the construction of field-level production profiles via kernel functions which are estimated nonparametrically using natural cubic splines. This flexible construction of production profiles makes it easy

to incorporate the assumption that the shapes of the production profiles vary continuously with the field's size (Michel (2011), p. 440).

Jakobsson et al. (2012) employ a field-level production model rooted in operations research which combines the geophysical facts of oil depletion with economic inter-temporal optimization of drilling and platform capacity additions. In this approach, the production profiles can be related to different scenarios for the oil price and cost parameters (see pp. 864-866). Then, Jakobsson et al. (2012) conduct a Monte Carlo simulation of size-biased sampling with the following assumptions: sampling of fields is exactly proportional to field size (recoverable oil), an exemplary lognormal distribution is assumed for the field sizes, and the drilling of unsuccessful exploration wells is endogenized by including an "empty area" that does not diminish with exploration. The simulation results are combined with three trends for the number of exploration wells per year and with the field-level production model, and the resulting regional discovery and production rates are shown in figures 12-13 of Jakobsson et al. (2012).

4.3 Mathematical Model

In a nutshell, the bottom-up model we formulate is composed of three elements. The first element are empirically founded field-level production profiles which convert the size of an oil field into a continuous time path for production. This is explained in sect. 4.3.1. The second element is a variant of the discovery process model from Barouch and Kaufman (1976a), which we refer to as the size-biased sampling model. We use this to predict the sequence of field sizes in order of discovery. In essence, the model allows the field sizes to decline (or increase) with the succession of fields in a stochastic fashion, and the degree to which this occurs can be estimated from the data. This is explained in sect. 4.3.2. As the third element, we assume that the discovery times of new fields are stochastic variables generated by a pure birth process. This is explained in sect. 4.3.3. Combining the three elements of the bottom-up model, a continuous time path for the rate of oil production in the region emerges simply as the sum of the individual field-level production paths.

Regarding notation, we stick to the notation of probability theory. Let $\mathbb{P}[\cdot]$ be the probability of a specified event, $\mathbb{E}[\cdot]$ be the expected value operation for scalar random variables (rvs) and random vectors, $\mathrm{Var}[\cdot]$ be the variance operation for scalar rvs, and $\mathbb{V}[\cdot]$ be the covariance matrix for random vectors. We denote random variables or vectors by upper case letters while we denote their possible realizations by the corresponding lower case letters. In case we deviate from this convention the definitions will be clearly stated in the text. Cumulative distribution functions are referred to as cdfs and probability density functions as pdfs. Besides, we occasionally make use of the Bayesian notation for pdfs (using the symbol $\propto$), which means that we leave out the normalizing constant. Also, for a continuous random vector with a density, say $f_{X,Y}(x, y)$, we often use the notation $\mathbb{P}[X \in \mathrm{d}x, Y \in \mathrm{d}y] = f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y$, which is common in probability theory. Also, we write $X \sim \cdot$ to express the assumption that X is distributed according to a certain probability distribution. Finally, let $\mathbb{1}(\cdot)$ denote the indicator function, $\perp\!\!\!\perp$ denote statistical independence, and $\top$ denote the transpose of a vector or matrix.

4.3.1 Overview and Production Profiles

In this section, we first formalize the aggregation from individual fields to the region, and then explain in detail how the field-level production profiles look like.

We define the stochastic process $(Q_\tau)_{\tau \geq 0}$ we are interested in as the superposition of the individual contributions to that process from all oil fields in the region. We write this formally as

$$Q_\tau \coloneqq \sum_{i=1}^{N} q(\tau, T_i, X_i). \tag{4.1}$$

Here, τ denotes real (physical) time, the index i labels the fields in order of discovery, and N denotes the number of fields in the region which are ultimately (economically or technically) recoverable. T_i and X_i are, respectively, discovery time and size of the i^{th} field, where we mean by size the ultimately recoverable resources (URR) of a field. The function q models the individual contribution that a field with a certain discovery time and size makes to the regional aggregate at time τ. As we are primarily interested in the modeling the rate of oil production, the function q should be a model for the production of an individual oil field. Aside from this, we introduce functions for q with which (4.1) becomes a model for remaining reserves or a "smoothed" discovery rate.

It is known empirically that for fields of similar size and physiographic situation (onshore, shallow offshore (shelf), deep water), oil production evolves quite similar as a function of time since first production (e.g., ch. 10 of IEA (2008); Höök et al., (2009b; 2009a); Höök (2014); Miller and Sorrell (2014)). This justifies using a deterministic function that depends only on the oil field's size and time *since* discovery (or first production), that is we can use a bivariate function $q_{\text{prod}}(\tau - t, x)$. In principle, any bivariate function is feasible if $\int_0^\infty q_{\text{prod}}(t, x)\mathrm{d}t = x$. Schematic figures of empirical profiles can be found in Höök et al. (2009b, p. 42) or Jakobsson (2014, p. 117). These profiles have a build-up phase of increasing oil production, followed by a plateau phase where new drilling or water-/gas-injection just compensates the pressure-driven decline, which is finally followed by a decline phase that is either exponential or hyperbolic (Höök (2014), p. 100).

Using a comprehensive data set, IEA (2008) reports in chapter 10 the results of a field-by-field analysis conducted to estimate standard production profiles for fields of similar size and physiographic situation. The results are presented by IEA (2008) via a few summary statistics, which we show in Table 4.1. In order to convert this into a fully parameterized production profile, we choose the function

$$q_{\text{prod}}(\tau - t, x) \coloneqq \begin{cases} 0 & \text{if } \tau - t \in (-\infty, \delta_0], \\ x \cdot \overline{q} \cdot \frac{(\tau - t - \delta_0)}{\delta_1 - \delta_0} & \text{if } \tau - t \in (\delta_0, \delta_1], \\ x \cdot \overline{q} \cdot \mathrm{e}^{-\vartheta_2(\tau - t - \delta_1)} & \text{if } \tau - t \in (\delta_1, \delta_2], \\ x \cdot 0.85\overline{q} \cdot \mathrm{e}^{-\vartheta_3(\tau - t - \delta_2)} & \text{if } \tau - t \in (\delta_2, \delta_3], \\ 0 & \text{if } \tau - t \in (\delta_3, \infty). \end{cases} \tag{4.2}$$

First, note that the function is linear in x, i.e. $q_{\text{prod}}(t, x) = q_{\text{prod}}(t, 1)x$. Nonlinear dependence on the field size will be achieved with different parameters for the different field-size categories shown in Table 4.1. The function implies that first oil is produced δ_0 years after the discovery. Production then starts with a linear build-up phase for the first $\delta_1 - \delta_0$ years, followed by a first phase of moderate exponential decline at rate ϑ_2 for the

Table 4.1: Summary statistics of standard production profiles from 725 oil fields analyzed by IEA (2008)

	% of initial reserves produced in peak year	% of initial reserves produced until peak year	Number of years of production at plateau[a]	Estimated average total number of production years[b]	Post-plateau decline rate
Onshore, <500 Mb	3.9	21	7	75	9.4%
Shelf, <500 Mb	9.7	25	4	60	12.2%
Onshore, 500 Mb - 1.5 Gb	2.3	17	10	90	5.5%
Shelf, 500 Mb - 1.5 Gb	3.5	20	8	65	8.6%
All, > 1.5 Gb	1.7	15	13	110	NA
Deepwater	7.0	22	5	27	12.6% (<500 Mb) 10.8% (>500 Mb)

Note: Mb=Mega barrels, Gb=Giga barrels

[a]Defined as the period during which production is more than 85% of that in the peak year.

[b]Over the full lifetime of the field, assuming that cumulative production strictly equals initial reserves

Source: taken from World Energy Outlook 2008, International Energy Agency, table 10.6 (p. 231), table 10.8 (p. 238)

next $\delta_2 - \delta_1$ years, after which we assume that production has reached 85% of the peak level. Thus, the two parameters ϑ_2 and $\delta_2 - \delta_1$ are related by $\exp(-\vartheta_2(\delta_2 - \delta_1)) = 0.85$. The profile then enters into the final phase of exponential decline at the higher rate ϑ_3 for $\delta_3 - \delta_2$ years. $\delta_3 - \delta_0$ is the time from first oil to abandonment, which occurs after production has declined below a threshold where revenues cannot compensate anymore for fixed operating cost.

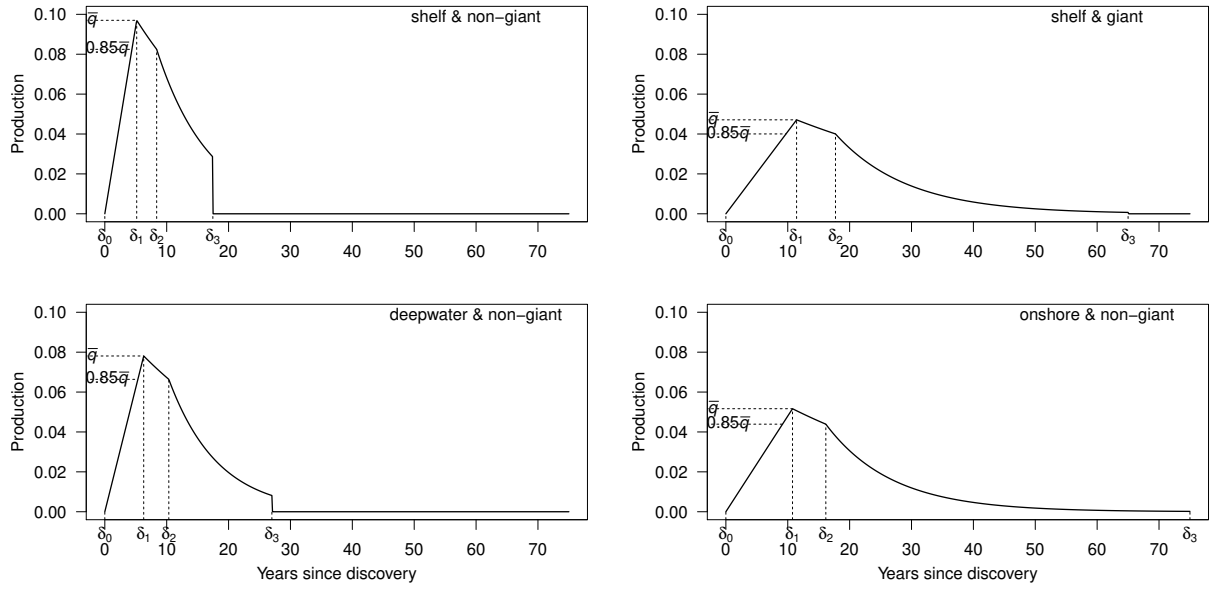
The first column of Table 4.1 is used as a value for the peak level, $\overline{q}$. From $\overline{q}$ together with the second and third columns, it is possible to determine the parameters $\delta_1 - \delta_0$, $\delta_2 - \delta_1$, and ϑ_2. For the decline rate parameter ϑ_3 we take values from table 10.8 of IEA (2008), where post-plateau decline rates are reported by size and physiographic situation. Unfortunately, decline rates for the size category >1.5 Gb are missing while only the category "super-giant" appears, which is defined by IEA (2008, p. 222) as >5 Gb. Therefore, we only classify fields into non-giants (<500 Mb) and giants (>500 Mb), using the numbers from the 500 Mb - 1.5 Gb categories for all giants.

Another problem is that when choosing $\overline{q}$ as in the first and $\delta_3 - \delta_0$ as in the fourth column, respectively, we find that the condition $\int_0^\infty q_{\text{prod}}(t, x)/x \mathrm{d}t = 1$ is not fulfilled exactly for the function (4.2). Therefore, for each category, we adjust the two parameters so that the condition holds exactly.[37] Figure 4.1 plots the resulting production profiles (4.2) for four different parameterizations from Table 4.1, with $\delta_0 = 0$ and a field size normalized to one.

Turning now to the parameter δ_0, the usual timeline between discovery and first oil production can be summarized as follows: after oil has been discovered by an exploration well, the surrounding area is geologically searched by further wells to assess the amount of oil contained in the field and the viability of recovering it (Mihalyi (2021), p. 7). This assessment phase is followed by feasibility studies with a detailed investigation of technical and financial aspects, and by the obtainment of necessary permits from the local regulator (Mihalyi (2021), p. 8). After the permits have been granted, the oil company can approve the development of the field. After approval, the necessary physical capital

[37]If the integral is greater than one, we fix $\overline{q}$ as in Table 4.1 and reduce $\delta_3 - \delta_0$ so that the condition is fulfilled. If the integral is less than one, we fix $\delta_3 - \delta_0$ and increase $\overline{q}$ so that the condition is fulfilled.

Figure 4.1: Four production profiles based on equation (4.2) and Table 4.1



Note: The scale of the y-axis is normalized so that each production profile integrates to one.

is procured and installed on-site, and only upon completion thereof the company can extract the first barrel of oil (Mihalyi (2021), pp. 8f.). The phase from discovery until approval usually takes several years and is usually longer than the capital procurement phase. Analyzing a comprehensive data set of about 20,500 oil and gas fields with first production before 2020, Mihalyi (2021) calculates a mean duration between discovery and approval of 5.5 years (with a standard deviation of 8.3 years), and a mean duration for the capital procurement phase of 1.5 years (with a standard deviation of 2.4 years).

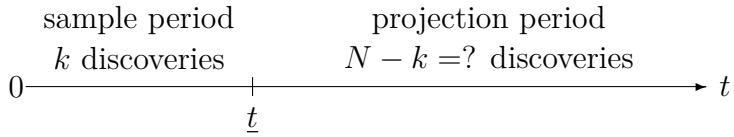
Mihalyi (2021) also analyzes the impact of geographical location and other variables on the duration between discovery and first production. The author finds strong differences by continent. While fields in the Americas have a mean duration of 6.4 years, fields in Sub-Saharan Africa have a mean duration of 16.8 years, and fields in Europe, Asia and other parts of Africa lie somewhere in between. The variability of the duration shows similar differences across the continents. Mihalyi (2021) also employs econometric duration analysis to examine the impact of field-level geological and country-level macroeconomic and institutional characteristics. Larger fields tend to have longer durations, although this finding seems to apply mainly to gas fields. As illustrated in figure 3 of Mihalyi (2021), for oil fields the durations increase only slightly with field size.

Once the production profile q_{prod} is fully specified, one can also substitute q-functions representing other quantities that are related to or derived from the production profile into (4.1), so that (4.1) describes the evolution over time of these related quantities aggregated to the region. Of particular interest can be the remaining reserves, which are defined as the difference between cumulative discoveries and cumulative production (Sorrell and Speirs (2009), p. 10). Thus, (4.1) can describe the regional remaining reserves by using

$$q_{\text{res}}(\tau - t, x) \coloneqq \mathbb{1}(\tau - t \geq 0) \cdot x - \int_0^{\tau - t} q_{\text{prod}}(u, x) \mathrm{d}u. \tag{4.3}$$

Moreover, a smoothed discovery rate at time τ can be obtained by using the rectangular

Figure 4.2: Overall timeline



kernel

$$q_{\text{disc}}(\tau - t, x) := \begin{cases} \frac{x}{\vartheta} & \text{if } |\tau - t| \leq \frac{\vartheta}{2}, \\ 0 & \text{else,} \end{cases} \tag{4.4}$$

where ϑ is a moving average window.

Finally, another q-function of interest may be a discounted value of the expected revenues or profits from producing a field according to a certain production profile. For this q-function the time parameter τ is integrated out by definition, so that (4.1) then does not define a stochastic process in τ but a single random variable, representing the discounted value of the expected revenue stream from the aggregate region. There are papers where such discounted production profiles are used. For example, Arezki et al. (2017) use production profiles to calculate present values of the revenues from fields in the giant oil field data set from Horn (2014), employing the oil price which was prevalent at the time of discovery and country-specific discount rates for valuation. However, it is out of the scope of this paper to combine such present value functions with the other components of the model presented here. This remains an interesting avenue for future research.

Let us end this section by clarifying the timeline of the overall exploration and production cycle in the region which we have in mind in the exposition of the following sections (see Figure 4.2). Time 0 is the date when first knowledge about significant oil deposits in the region is obtained so that commercial exploration ensues, and time $\underline{t}$ denotes "now", after $k \geq 0$ discoveries have been made and from where one aims to forecast future oil discovery and production.

4.3.2 Size-Biased Sampling Model

In this section, we present the assumptions and parameter estimation strategy for the size-biased sampling model. Due to the mathematical complexity of the model, we use several result not proven here, where we only cite the relevant literature. In order that this paper is self-contained, we complement this section with our own derivations of some relevant results that are known, and also of some new results. This can be found in sects. B.1-B.5 in the Appendix.

A parametric discovery process model usually starts with an assumption about the field-size distribution, also referred to as the superpopulation distribution. This can be stated as follows (e.g., Kaufman et al. (1975), p. 15; Barouch and Kaufman (1976a), p. 2; Nair and Wang (1989), p. 423; Lee (2008), pp. 26ff.):

Assumption 1. *Nature has generated the oil fields with sizes $\{\tilde{X}_1, \ldots, \tilde{X}_n\}$, where the relative frequencies of the sizes follow some probability law which can be described by a pdf $f_{\boldsymbol{\theta}}$ that is known up to an unknown vector of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Letting the indices*

$\{1,\ldots,n\}$ *be uninformative about the sizes, this data generating process is mathematically a random sample* $\tilde{\mathcal{X}}_n \coloneqq (\tilde{X}_1,\ldots,\tilde{X}_n)$ *from* $f_{\boldsymbol{\theta}}$.

Already by Arps and Roberts (1958) and Kaufman (1963), oil field sizes and related geological random variables have been approximated by a lognormal distribution, since empirically these variables show distributions that are positively skewed with long right tails (Kaufman et al. (1975), p. 15). The lognormal, Pareto, and Weibull distributions are the most common size distributions with these characteristics (Lee (2008), pp. 61ff.; Jakobsson et al. (2014), p. 120). In this paper, we assume that oil field sizes are generated from a (truncated) distribution from the regular exponential family, which includes the Pareto and lognormal distribution, but excludes the Weibull distribution. We restrict the class of distributions to the exponential family since this is convenient from a statistical point for estimating the parameters as we will see later. We also assume that $f_{\boldsymbol{\theta}}$ is truncated with support $[\underline{x},\overline{x}]$, and use the notation $f_{\boldsymbol{\theta}}/p_{\text{tr}}(\boldsymbol{\theta})$ from here on, where $p_{\text{tr}}(\boldsymbol{\theta}) \coloneqq \int_{\underline{x}}^{\overline{x}} f_{\boldsymbol{\theta}}(x)\mathrm{d}x$. The support is assumed to be known or estimated on a rule of thumb basis from the observed field sizes.

A second assumption is added which lends the name to the size-biased sampling model. This can be stated as follows (e.g., Nair and Wang (1989), p. 423; Lee (2008), p. 30):

Assumption 2. *Given that the oil fields which exist in the region have the sizes* $\tilde{\mathcal{X}}_n = (\tilde{X}_1,\ldots,\tilde{X}_n)$*, and assuming that all fields are in principle available for discovery at each time, the order of discovery is sampled from* $\tilde{\mathcal{X}}_n$ *without replacement and with probabilities proportional to some positive, monotonous transformation of the size,* $w(x)$*. For oil discovery, this function is usually chosen as* $w(x) = x^{\beta}$*, where* β *is called the discoverability parameter.*[38]

To achieve a concise formal notation, we introduce the outcome sets for permutations and drawing without replacement from the letters $\{1,\ldots,n\}$ as

$$\begin{aligned}\mathcal{I}_n &\coloneqq \big\{(i(1),\ldots,i(n)) \in \{1,\ldots,n\}^n \big| i(r) \neq i(s) \text{ for } r \neq s\big\},\\ \mathcal{I}_{l,n} &\coloneqq \big\{(i(1),\ldots,i(l)) \in \{1,\ldots,n\}^l \big| i(r) \neq i(s) \text{ for } r \neq s\big\}.\end{aligned}$$

The successive sampling from $\tilde{\mathcal{X}}_n$ can be described formally by a random vector that is a permutation of the letters $\{1,\ldots,n\}$. Let $(\sigma(1),\ldots,\sigma(n))$ denote this random vector, containing the n indices which relate the field sizes as labeled in the order of discovery, X_i, to the uninformatively labeled field sizes, $\tilde{X}_i$, that is $X_1 \coloneqq \tilde{X}_{\sigma(1)},\ldots,X_n \coloneqq \tilde{X}_{\sigma(n)}$. Expressed formally, assumption 2 implies that the conditional probability mass function (pmf) of observing $X_l = \tilde{X}_{i(l)}$, for any $l \leq n$ and any distinct indices $(i(1),\ldots,i(l)) \in \mathcal{I}_{l,n}$, equals (Nair and Wang (1989), p. 423; Pitman and Tran (2015), p. 2484)

$$\mathbb{P}\big[\sigma(l) = i(l)\big|\sigma(1) = i(1),\ldots,\sigma(l-1) = i(l-1),\tilde{\mathcal{X}}_n\big] = \frac{\tilde{X}_{i(l)}^{\beta}}{S_n - \tilde{X}_{i(1)}^{\beta} - \ldots - \tilde{X}_{i(l-1)}^{\beta}}, \tag{4.5}$$

where $S_n \coloneqq \tilde{X}_1^{\beta} + \ldots + \tilde{X}_n^{\beta}$.

[38] The reason why $w(x) = x^{\beta}$ is convenient is that it nests several sampling-schemes. For $\beta = 0$, the density of observing the first k sizes $(x_1,....,x_k)$ reduces to that of iid sampling. For $\beta = 1$, the sampling becomes exactly proportional to size. For $\beta \to \infty$ $(-\infty)$, it is certain that the largest (smallest) field will be discovered first, then the second largest (smallest), and so on. Thus, the parameterization also nests the order statistics case.

Multiplying over $l = 1, ..., n$ yields the pmf of observing $X_1 = \tilde{X}_{i(1)}, \ldots, X_n = \tilde{X}_{i(n)}$ for given values $\tilde{\mathcal{X}}_n = (\tilde{X}_1, \ldots, \tilde{X}_n)$ and any index permutation $(i(1), \ldots, i(n)) \in \mathcal{I}_n$,

$$\begin{aligned}\mathbb{P}\big[\sigma(1) = i(1), \ldots, \sigma(n) = i(n) \big| \tilde{\mathcal{X}}_n\big] &= \frac{\tilde{X}_{i(1)}^{\beta}}{S_n} \cdot \frac{\tilde{X}_{i(2)}^{\beta}}{S_n - \tilde{X}_{i(1)}^{\beta}} \cdot \ldots \cdot \frac{\tilde{X}_{i(n)}^{\beta}}{S_n - \tilde{X}_{i(1)}^{\beta} - \ldots - \tilde{X}_{i(n-1)}^{\beta}} \\ &= \frac{\tilde{X}_{i(1)}^{\beta}}{\tilde{X}_{i(1)}^{\beta} + \ldots + \tilde{X}_{i(n)}^{\beta}} \cdot \frac{\tilde{X}_{i(2)}^{\beta}}{\tilde{X}_{i(2)}^{\beta} + \ldots + \tilde{X}_{i(n)}^{\beta}} \cdot \ldots \cdot \frac{\tilde{X}_{i(n)}^{\beta}}{\tilde{X}_{i(n)}^{\beta}},\end{aligned} \tag{4.6}$$

where the second equality follows from rewriting the denominator. Combining this with assumption 1 yields the joint distribution of $(\sigma(1), \ldots, \sigma(n), X_1, \ldots, X_n)$, from which the distribution of $(X_1, \ldots, X_n)$ is readily derived as follows

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}}\big[\sigma(1) = i(1), \ldots, \sigma(n) = i(n), X_1 \in \mathrm{d}x_1, \ldots, X_n \in \mathrm{d}x_n\big] \\ &= \mathbb{P}_{\boldsymbol{\theta}}\big[\sigma(1) = i(1), \ldots, \sigma(n) = i(n), \tilde{X}_{i(1)} \in \mathrm{d}x_1, \ldots, \tilde{X}_{i(n)} \in \mathrm{d}x_n\big] \\ &= \Big(\prod_{j=1}^{n} \frac{f_{\boldsymbol{\theta}}(x_j)}{p_{\mathrm{tr}}(\boldsymbol{\theta})} \mathrm{d}x_j\Big) \cdot \Big(\prod_{j=1}^{n} \frac{x_j^{\beta}}{x_j^{\beta} + \ldots + x_n^{\beta}}\Big), \\ \Rightarrow &\mathbb{P}_{\boldsymbol{\theta}}[X_1 \in \mathrm{d}x_1, \ldots, X_n \in \mathrm{d}x_n] = n! \prod_{j=1}^{n} \frac{f_{\boldsymbol{\theta}}(x_j)}{p_{\mathrm{tr}}(\boldsymbol{\theta})} \frac{x_j^{\beta}}{x_j^{\beta} + \ldots + x_n^{\beta}} \mathrm{d}x_j.\end{aligned} \tag{4.7}$$

The last line is due to the exchangeability of the iid rvs $(\tilde{X}_1, \ldots, \tilde{X}_n)$, which makes the joint distribution evaluated at each index permutation $(i(1), \ldots, i(n)) \in \mathcal{I}_n$ identical, and due to $|\mathcal{I}_n| = n!$.

Let us now consider only the first k field sizes, which we collect into the vector $\boldsymbol{X} \coloneqq (X_1, \ldots, X_k)$. By analogous derivations a formula for the distribution of $\boldsymbol{X}$ can be derived. Upon introducing the notation $b_j \coloneqq x_j^{\beta} + \ldots + x_k^{\beta}$ for $j = 1, \ldots, k$, and letting S_{n-k} be the sum of $n - k$ iid rvs which are distributed like $\tilde{X}_1^{\beta}$, it holds that (Barouch and Kaufman (1976a), p. 3; Pitman and Tran (2015), p. 2486)

$$\mathbb{P}_{\boldsymbol{\theta}}[X_1 \in \mathrm{d}x_1, \ldots, X_k \in \mathrm{d}x_k] = \frac{n!}{(n-k)!} \Big(\prod_{j=1}^{k} \frac{f_{\boldsymbol{\theta}}(x_j)}{p_{\mathrm{tr}}(\boldsymbol{\theta})} \frac{x_j^{\beta}}{b_j} \mathrm{d}x_j\Big) \cdot \mathbb{E}_{\boldsymbol{\theta}}\Big[\prod_{j=1}^{k} \frac{b_j}{b_j + S_{n-k}}\Big]. \tag{4.8}$$

In the context of oil discovery, it is reasonable to add a third probabilistic assumption about the total number of fields in the region which is so far regarded as a fixed quantity. Lee (2008, p. 38, pp. 195f.) proposes to approximate the number of fields with a Poisson prior distribution in the absence of detailed geological knowledge that would specify a different distribution. This effectively replaces the fixed quantity n by a parameter for the expected number of fields in the region, which we denote by ν. A Poisson distribution for the number of oil accumulations might also be justified on the basis of a spatial Poisson point process, which is a common model in mineral prospectivity analysis (e.g. Baddeley (2018)).

Assumption 3. *The total number of fields in the region has a Poisson prior:* $N \sim \mathrm{Pois}(\nu)$.

Let us consider the information available to a forecaster at time $\underline{t}$ (see Figure 4.2 for the time arrow). The forecaster observes that the first k field sizes in order of discovery assume the values $x_1, \ldots, x_k$, and infers from this that the region contains a total of $N \geq k$

fields. We refer to this information as the *observed data*, and summarize it formally by $\mathcal{F} \coloneqq \big\{\boldsymbol{X} = (x_1, \ldots, x_k), N \geq k\big\}$ (for $k > 0$).

From the Poisson prior and the likelihood function (eq. 4.8, to be arranged in a different form in eq. 4.13), we derive a mixed Poisson distribution as the *Bayesian posterior* of $N-k$, the number of undiscovered fields, after the information in $\mathcal{F}$ has been conditioned on. We prove this in Lemma 2 in Appendix B.1. The final result is that this posterior is the distribution $(N-k)|U_k \sim \text{Pois}(\nu U_k)$, where U_k is a mixing rv that is data-dependent (i.e. dependent on $(x_1, \ldots, x_k)$), and assumes values in $(0, 1]$.[39] Although the formula for the posterior mean of $N-k$ under a Poisson prior has been derived by Lee (2008, pp. 195f.), the fact that the posterior distribution is a (mixed) Poisson distribution has not been established in this reference, and to the best of our knowledge, it has not been stated somewhere else in the discovery process model literature.

After knowing the distribution of the number of undiscovered fields, we also need the distribution of the sizes of these fields. Reconsidering the likelihood function (4.8), one can see that these sizes are contained there in the sum S_{n-k} with respect to which the expectation is taken. Conditional on $N = n$, the *joint distribution* of the k *discovered* and the $n-k$ *undiscovered* field sizes is obtained from (4.8) by writing out the expectation over $S_{n-k} = \tilde{X}_{k+1}^{\beta} + \ldots + \tilde{X}_{n}^{\beta}$, where the $\tilde{X}_i$ are mutually iid with pdf $f_{\boldsymbol{\theta}}/p_{\text{tr}}(\boldsymbol{\theta})$ as defined above, and by dropping the integral with respect to these variables. To express this as a joint density function we write $\tilde{\boldsymbol{X}} \coloneqq (\tilde{X}_{k+1}, \ldots, \tilde{X}_n)$ for the random vector of the undiscovered field sizes, and $\tilde{\boldsymbol{x}} \coloneqq (\tilde{x}_{k+1}, \ldots, \tilde{x}_n)$ for a possible realization. We also refer to the collections $(\tilde{\boldsymbol{X}}, \boldsymbol{X})|N$ (under assumptions 1 - 2 alone) and $(\tilde{\boldsymbol{X}}, \boldsymbol{X}, N)$ (under assumptions 1 - 3) as the *complete data*. Thus, under assumptions 1 - 2 and under assumptions 1 - 3, respectively, the pdf of the *complete data* equals

$$
\begin{aligned}
f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}, \boldsymbol{x}|n) &= \frac{n!}{(n-k)!}\Big(\prod_{i=1}^{k} \frac{f_{\boldsymbol{\theta}}(x_i)}{p_{\text{tr}}(\boldsymbol{\theta})} \frac{x_i^{\beta}}{b_i}\Big)\Big(\prod_{i=1}^{k} \frac{b_i}{b_i + \tilde{x}_{k+1}^{\beta} + \ldots + \tilde{x}_n^{\beta}} \prod_{i=k+1}^{n} \frac{f_{\boldsymbol{\theta}}(\tilde{x}_i)}{p_{\text{tr}}(\boldsymbol{\theta})}\Big), \\
f_{(\boldsymbol{\theta},\nu)}(\tilde{\boldsymbol{x}}, \boldsymbol{x}, n) &= \mathrm{e}^{-\nu} \frac{\nu^n}{(n-k)!}\Big(\prod_{i=1}^{k} \frac{f_{\boldsymbol{\theta}}(x_i)}{p_{\text{tr}}(\boldsymbol{\theta})} \frac{x_i^{\beta}}{b_i}\Big)\Big(\prod_{i=1}^{k} \frac{b_i}{b_i + \tilde{x}_{k+1}^{\beta} + \ldots + \tilde{x}_n^{\beta}} \prod_{i=k+1}^{n} \frac{f_{\boldsymbol{\theta}}(\tilde{x}_i)}{p_{\text{tr}}(\boldsymbol{\theta})}\Big).
\end{aligned}
\tag{4.9}
$$

Essentially, the preceding discussion implies for statistical inference that this is an incomplete data problem where one observes only a subset of the complete data, and where the observed and unobserved data are jointly distributed in a non-independent fashion, as (4.9) shows. For these type of problems, the *expectation-maximization* (EM) algorithm as introduced by Dempster et al. (1977) can be applied to find the maximum likelihood estimator of the observed data likelihood by iteratively maximizing the expected log-likelihood of the complete data. The use of the EM algorithm for estimating the parameters of discovery process models was first proposed by Barouch et al. (1983).

As described by Dempster et al. (1977, pp. 6ff.) in its general form, starting from an

[39] This result is best verified by the reader by first continuing to read until after (4.15), where U_k is defined directly as a function of another mixing rv Γ_k, whose distribution is also specified there, and then going through the proof of Lemma 2 in in Appendix B.1.

initial guess $\boldsymbol{\theta}_0$, each iteration of the EM algorithm can be summarized as follows:

EM iteration $\boldsymbol{\theta}_r \to \boldsymbol{\theta}_{r+1}$:

E-step: Compute the expectation of the complete data log-likelihood as a function of $\boldsymbol{\theta}' \in \Theta$, where the expectation is taken with respect to the distribution of the unobserved given the observed data, which is parametrized by $\boldsymbol{\theta} = \boldsymbol{\theta}_r$.

M-step: Choose $\boldsymbol{\theta}_{r+1}$ as a value $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ that maximizes this function.

The remainder of this section is concerned with deriving under assumptions 1 - 3 the EM iterations from the complete data pdf and the Hessian matrix of the observed data log-likelihood. The final result of this parameter estimation strategy is summarized in Theorem 1.

To enable the necessary derivations, we first introduce a well-known integral representation of the expectation in (4.8), and state how this formula changes by adding assumption 3. It has been shown in the literature (e.g., Barouch and Kaufman (1976a), pp. 12f.; Nair and Wang (1989), p. 427; Lee (2008), p. 179) that the expectation or integral,

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\prod_{i=1}^{k}\frac{b_i}{b_i+\tilde{X}_{k+1}^{\beta}+...+\tilde{X}_{n}^{\beta}}\Big] = \int\limits_{[\underline{x},\overline{x}]^{n-k}}\prod_{i=1}^{k}\frac{b_i}{b_i+\tilde{x}_{k+1}^{\beta}+...+\tilde{x}_{n}^{\beta}}\prod_{i=k+1}^{n}\frac{f_{\boldsymbol{\theta}}(\tilde{x}_i)}{p_{\mathrm{tr}}(\boldsymbol{\theta})}\mathrm{d}\tilde{x}_{k+1}\cdots\mathrm{d}\tilde{x}_n, \tag{4.10}$$

can be rearranged by several nontrivial steps into a one-dimensional integral that includes in the integrand not anymore the fractions in (4.10) but instead powers of the *Laplace transform of* $\tilde{X}_1^{\beta}$. This *Laplace transform* is defined by

$$\phi_{\boldsymbol{\theta}}(\gamma) \coloneqq \mathbb{E}_{\boldsymbol{\theta}}\Big[\mathrm{e}^{-\gamma\tilde{X}_1^{\beta}}\Big] = \int_{\underline{x}}^{\overline{x}}\mathrm{e}^{-\gamma x^{\beta}}\frac{f_{\boldsymbol{\theta}}(x)}{p_{\mathrm{tr}}(\boldsymbol{\theta})}\mathrm{d}x. \tag{4.11}$$

As a second term the integrand includes a data-dependent density function. In particular, this is the convolution of k exponential pdfs with parameters $b_1,\ldots,b_k$,[40] which is why it is referred to as the *general gamma* (or Erlang) density (McGill and Gibbon (1965), pp. 4f.; Nair and Wang (1989), p. 427). We denote this density by $g_{\boldsymbol{b}}(\gamma)$ since it is parameterized by the vector $\boldsymbol{b} \coloneqq (b_1,\ldots,b_k)$. It is sensitive to the discoverability parameter β *and* the order in which the field sizes $(x_1,\ldots,x_k)$ are observed. With these definitions, the alternative integral representation of (4.10) is as follows

$$\begin{aligned}\prod_{i=1}^{k}\frac{b_i}{b_i+\tilde{x}_{k+1}^{\beta}+...+\tilde{x}_{n}^{\beta}} &= \int_0^{\infty}\Big(\prod_{i=k+1}^{n}\mathrm{e}^{-\gamma\tilde{x}_i^{\beta}}\Big)g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma,\\ \Rightarrow \mathbb{E}_{\boldsymbol{\theta}}\Big[\prod_{i=1}^{k}\frac{b_i}{b_i+\tilde{X}_{k+1}^{\beta}+...+\tilde{X}_{n}^{\beta}}\Big] &= \int_0^{\infty}\phi_{\boldsymbol{\theta}}(\gamma)^{n-k}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma,\end{aligned} \tag{4.12}$$

where the interchange of the order of integration in the second line is allowed by Fubini's theorem because all terms are positive.

A complication of this approach, though, is that the general gamma density does not have a useful analytic expression. However, there is a feasible numerical approach employed in the literature, the *Fourier-series method* for inverting the Laplace transform of the

[40] A convolution emerges in probability theory when summing independent rvs, where the sum then has a pdf equal to the convolution of the individual pdfs.

density. We describe the details of this method and how we implement it in Appendix B.2.

Applying (4.12) to (4.8) rearranges the pdf of the observed data under assumptions 1 - 2. By adding assumption 3 we also need to take the mean of the resulting expression with respect to the Poisson prior, which does not further complicate the expression (we derive the formula in Lemma 1 in Appendix B.1). Denoting the pdf or likelihood of the observed data by ℓ, this function becomes under assumptions 1 - 2 and under assumptions 1 - 3, respectively

$$\begin{aligned}
\ell_{\boldsymbol{\theta}}(\boldsymbol{x}|n) &= \frac{n!}{(n-k)!}\Big(\prod_{i=1}^{k}\frac{f_{\boldsymbol{\theta}}(x_i)}{p_{\mathrm{tr}}(\boldsymbol{\theta})}\frac{x_i^{\beta}}{b_i}\Big)\Big(\int_0^{\infty}\phi_{\boldsymbol{\theta}}(\gamma)^{n-k}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\Big),\\
\ell_{(\boldsymbol{\theta},\nu)}(\mathcal{F}) &= \nu^k\Big(\prod_{i=1}^{k}\frac{f_{\boldsymbol{\theta}}(x_i)}{p_{\mathrm{tr}}(\boldsymbol{\theta})}\frac{x_i^{\beta}}{b_i}\Big)\Big(\int_0^{\infty}\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\Big),\\
\Rightarrow L(\boldsymbol{\theta},\nu) &\coloneqq \ln(\ell_{(\boldsymbol{\theta},\nu)}(\mathcal{F})).
\end{aligned} \tag{4.13}$$

From (4.13) we obtain $L(\boldsymbol{\theta},\nu)$ as the log-likelihood we want to maximize.

Next, what is required to derive the EM iterations is the posterior distribution of the unobserved given the observed data. Upon introducing a new mixing rv $\mathit{\Gamma}_k$, the collections $(\tilde{\boldsymbol{X}},\mathit{\Gamma}_k)|(\boldsymbol{X},N)$ and $(\tilde{\boldsymbol{X}},N,\mathit{\Gamma}_k)|\mathcal{F}$, respectively, can be shown to have pdfs

$$\begin{aligned}
\rho_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}},\gamma|\boldsymbol{x},n) &= \Big(\prod_{i=k+1}^{n}\frac{\mathrm{e}^{-\gamma\tilde{x}_i^{\beta}}f_{\boldsymbol{\theta}}(\tilde{x}_i)}{\phi_{\boldsymbol{\theta}}(\gamma)p_{\mathrm{tr}}(\boldsymbol{\theta})}\Big)\cdot\Big(\frac{\phi_{\boldsymbol{\theta}}(\gamma)^{n-k}g_{\boldsymbol{b}}(\gamma)}{\int_0^{\infty}\phi_{\boldsymbol{\theta}}(\gamma)^{n-k}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}\Big),\\
\rho_{(\boldsymbol{\theta},\nu)}(\tilde{\boldsymbol{x}},n,\gamma|\mathcal{F}) &= \Big(\prod_{i=k+1}^{n}\frac{\mathrm{e}^{-\gamma\tilde{x}_i^{\beta}}f_{\boldsymbol{\theta}}(\tilde{x}_i)}{\phi_{\boldsymbol{\theta}}(\gamma)p_{\mathrm{tr}}(\boldsymbol{\theta})}\Big)\cdot\Big(\frac{(\nu\phi_{\boldsymbol{\theta}}(\gamma))^{n-k}}{(n-k)!}\mathrm{e}^{-\nu\phi_{\boldsymbol{\theta}}(\gamma)}\Big)\\
&\quad\cdot\Big(\frac{\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)}{\int_0^{\infty}\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}\Big).
\end{aligned} \tag{4.14}$$

To allow for a more compact notation we also define the two univariate pdfs

$$\begin{aligned}
\rho_{\boldsymbol{\theta}}(\tilde{x}|\gamma) &\coloneqq \frac{\mathrm{e}^{-\gamma\tilde{x}^{\beta}}f_{\boldsymbol{\theta}}(\tilde{x})}{\phi_{\boldsymbol{\theta}}(\gamma)p_{\mathrm{tr}}(\boldsymbol{\theta})},\\
g^{(0)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F}) &\coloneqq \frac{\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)}{\int_0^{\infty}\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}.
\end{aligned} \tag{4.15}$$

The first density in (4.14) can be derived by dividing $f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}},\boldsymbol{x}|n)$ by $\ell_{\boldsymbol{\theta}}(\boldsymbol{x}|n)$, as shown in Nair and Wang (1989, p. 427) or Lee (2008, p. 184). We prove in Lemma 2 in Appendix B.1 how the Poisson prior assumption allows to derive the second density in (4.14) from the first density. Based on this, we define the mixing rv U_k from the discussion below assumption 3 as $U_k \coloneqq \phi_{\boldsymbol{\theta}}(\mathit{\Gamma}_k)$, where $\mathit{\Gamma}_k$ has the pdf $g^{(0)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})$. The second line in (4.14) contains the Poisson posterior result that $(N-k)|U_k \sim \mathrm{Pois}(\nu U_k)$. Besides, the first term in (4.14) makes it clear that $\tilde{\boldsymbol{X}}$ is distributed as a mixture of iid rvs: for fixed values of (n,γ) it contains $n-k$ iid elements from pdf $\rho_{\boldsymbol{\theta}}(\tilde{x}|\gamma)$.

Finally, we alter the complete data likelihood from (4.9) so that the truncation factors cancel and it becomes easier to maximize via the EM algorithm. This general approach was also introduced by Dempster et al. (1977, pp. 13f.) and it involves "artificially" extending the complete data vector by certain carefully constructed random variables, which does not change the observed data likelihood, but cancels the $p_{\mathrm{tr}}(\boldsymbol{\theta})$-terms in the complete

data likelihood. Specifically, let $M|N \sim \text{NegBin}(N, p_{\text{tr}}(\boldsymbol{\theta}))$ (negative binomial distribution), and let $\tilde{\boldsymbol{Y}} \coloneqq (\tilde{Y}_1, \ldots, \tilde{Y}_M)$ be a random sample from pdf $f_{\boldsymbol{\theta}}(\tilde{y})/\big(1 - p_{\text{tr}}(\boldsymbol{\theta})\big)$ with support on $(0, \underline{x}) \bigcup (\overline{x}, \infty)$, where each $\tilde{Y}_i$ is independent of $(M, \tilde{\boldsymbol{X}}, \boldsymbol{X}, N)$.[41] Multiplying (4.9) by the conditional pdf of $(\tilde{\boldsymbol{Y}}, M)|N$, the joint pdf of $(\tilde{\boldsymbol{Y}}, M, \tilde{\boldsymbol{X}}, \boldsymbol{X}, N)$ becomes

$$\begin{aligned} f_{(\boldsymbol{\theta},\nu)}(\tilde{\boldsymbol{y}}, m, \tilde{\boldsymbol{x}}, \boldsymbol{x}, n) = \mathrm{e}^{-\nu}\nu^n \prod_{i=1}^{k} f_{\boldsymbol{\theta}}(x_i) \prod_{i=1}^{m} f_{\boldsymbol{\theta}}(\tilde{y}_i) \prod_{i=k+1}^{n} f_{\boldsymbol{\theta}}(\tilde{x}_i) \\ \cdot \left\{ \frac{1}{(n-k)!} \binom{m+n-1}{m} \prod_{i=1}^{k} \frac{x_i^{\beta}}{b_i + \tilde{x}_{k+1}^{\beta} + \ldots + \tilde{x}_n^{\beta}} \right\}. \end{aligned} \tag{4.16}$$

Denoting the expected complete data log-likelihood derivable from (4.16) by $F(\boldsymbol{\theta}', \nu'|\boldsymbol{\theta}, \nu)$, this function equals[42]

$$\begin{aligned} F(\boldsymbol{\theta}', \nu'|\boldsymbol{\theta}, \nu) &= \mathbb{E}_{(\boldsymbol{\theta},\nu)}\Big[\ln f_{(\boldsymbol{\theta}',\nu')}(\tilde{\boldsymbol{Y}}, M, \tilde{\boldsymbol{X}}, \boldsymbol{X}, N)\Big|\mathcal{F}\Big] \\ &= \sum_{i=1}^{k} \ln f_{\boldsymbol{\theta}'}(x_i) + \mathbb{E}_{(\boldsymbol{\theta},\nu)}\Big[\sum_{i=1}^{M} \ln f_{\boldsymbol{\theta}'}(\tilde{Y}_i)\Big|\mathcal{F}\Big] + \mathbb{E}_{(\boldsymbol{\theta},\nu)}\Big[\sum_{i=k+1}^{N} \ln f_{\boldsymbol{\theta}'}(\tilde{X}_i)\Big|\mathcal{F}\Big] \\ &\quad - \nu' + \ln(\nu')\mathbb{E}_{(\boldsymbol{\theta},\nu)}\big[N\big|\mathcal{F}\big] + \text{constant}. \end{aligned} \tag{4.17}$$

In the EM framework it is useful to restrict $f_{\boldsymbol{\theta}}$ to be a member of the regular exponential family, where any distribution from this class can be expressed as $f_{\boldsymbol{\eta}}(x) = c(x)\mathrm{e}^{\boldsymbol{\eta}^{\top}\boldsymbol{s}(x)}/a(\boldsymbol{\eta})$. Here, the so-called "natural" parameters $\boldsymbol{\eta}$ are a vector-valued transformation of $\boldsymbol{\theta}$, and the vector-valued function $\boldsymbol{s}(x)$ contains the corresponding sufficient statistics. The univariate function $c(x)$ depends only on the observation and $a(\boldsymbol{\eta})$ only on the parameters. For distributions from the regular exponential family the EM iterations as outlined above simplify considerably. In particular there is no need to compute the expectation of the complete data log-likelihood, it is only necessary to compute the expectations of the sufficient statistics. The exponential family form of the lognormal distribution is summarized in Table B1 in Appendix B.3.

The topic of Theorem 1 is now to summarize the EM iterations derivable from (4.17) and the Hessian derivable from (4.13) for (truncated) distributions from the regular exponential family, such as the lognormal or Pareto distribution. The EM iteration in part (1) of the Theorem (without the term to adjust for the truncated distribution) appears in Lee (2008, p. 196). To our knowledge, previous derivations of the Hessian matrix have only been done for the model based on assumptions 1-2 only. The Hessian matrix in part (2) of the Theorem is thus a novel extension.

Theorem 1.
(1). For any (truncated) distribution $f_{\boldsymbol{\theta}}/p_{\text{tr}}(\boldsymbol{\theta})$ from the regular exponential family, the EM iteration $(\boldsymbol{\theta}_r, \nu_r) \to (\boldsymbol{\theta}_{r+1}, \nu_{r+1})$ under assumptions 1 - 3 is obtained by solving the

[41] Because the random variable M is used only in this section there should be no confusion with the stochastic birth process $(M_u)_{u\geq 0}$ introduced in sect. 4.3.3.

[42] The "constant" accounts for the curly bracket of (4.16) which does not depend on the parameters $(\boldsymbol{\theta}', \nu')$ and thus is irrelevant for maximization.

following expected moment equations[43] [44]

$$\begin{aligned}\frac{\partial \ln a(\boldsymbol{\theta}_{r+1})}{\partial \boldsymbol{\eta}} &= p_{\mathrm{tr}}(\boldsymbol{\theta}_r)\alpha(\boldsymbol{\theta}_r,\nu_r)\Big(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{s}(x_i)\Big) + \big(1-p_{\mathrm{tr}}(\boldsymbol{\theta}_r)\big)\int\limits_{(0,\underline{x})\cup(\overline{x},\infty)} \boldsymbol{s}(x)\frac{f_{\boldsymbol{\theta}_r}(x)}{1-p_{\mathrm{tr}}(\boldsymbol{\theta}_r)}\mathrm{d}x \\ &\quad + p_{\mathrm{tr}}(\boldsymbol{\theta}_r)\big(1-\alpha(\boldsymbol{\theta}_r,\nu_r)\big)\int_0^{\infty}\int_{\underline{x}}^{\overline{x}} \boldsymbol{s}(x)\rho_{\boldsymbol{\theta}_r}(x|\gamma)g^{(1)}_{(\boldsymbol{\theta}_r,\nu_r)}(\gamma|\mathcal{F})\mathrm{d}x\mathrm{d}\gamma, \\ \nu_{r+1} &= k + \mathbb{E}_{(\boldsymbol{\theta}_r,\nu_r)}[N-k|\mathcal{F}],\end{aligned}$$

where $\alpha(\boldsymbol{\theta},\nu) = k\big/\big(k+\mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}]\big)$*, and where* $g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})$ *and* $\mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}]$ *are*

$$\begin{aligned} g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F}) &\propto \nu\phi_{\boldsymbol{\theta}}(\gamma)\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma), \\ \mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}] &= \frac{\int_0^{\infty}\nu\phi_{\boldsymbol{\theta}}(\gamma)\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}{\int_0^{\infty}\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}. \end{aligned} \tag{4.18}$$

(2). For any (truncated) distribution $f_{\boldsymbol{\theta}}/p_{\mathrm{tr}}(\boldsymbol{\theta})$ *from the regular exponential family, the Hessian matrix of the log-likelihood under assumptions 1 - 3 is given by*

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\theta},\nu)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\top}} &= -\,k\cdot\mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big] \\ &\quad +\ \mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}]\cdot\Big\{\int_0^{\infty}\Big(\mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\} \\ &\quad +\ \mathbb{E}_{(\boldsymbol{\theta},\nu)}[(N-k)^2|\mathcal{F}]\cdot\Big\{\int_0^{\infty}\Big(\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big) \\ &\qquad\qquad \cdot\Big(\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)^{\top}g^{(2)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\} \\ &\quad -\ \big(\mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}]\big)^2\cdot\Big\{\int_0^{\infty}\Big(\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\} \\ &\qquad\qquad \cdot\Big\{\int_0^{\infty}\Big(\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)^{\top}g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\}.\end{aligned}$$

Here, the conditional mean vector and covariance matrix are taken with respect to the conditional pdf $\rho_{\boldsymbol{\theta}}(x|\gamma)$*, the unconditional mean vector and covariance matrix are taken with respect to the pdf* $f_{\boldsymbol{\theta}}(x)/p_{\mathrm{tr}}(\boldsymbol{\theta})$*, and* $g^{(2)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})$ *and* $\mathbb{E}_{(\boldsymbol{\theta},\nu)}[(N-k)^2\big|\mathcal{F}]$ *are*

$$\begin{aligned} g^{(2)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F}) &\propto \nu\phi_{\boldsymbol{\theta}}(\gamma)\big(1+\nu\phi_{\boldsymbol{\theta}}(\gamma)\big)\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma), \\ \mathbb{E}_{(\boldsymbol{\theta},\nu)}[(N-k)^2\big|\mathcal{F}] &= \frac{\int_0^{\infty}\nu\phi_{\boldsymbol{\theta}}(\gamma)\big(1+\nu\phi_{\boldsymbol{\theta}}(\gamma)\big)\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}{\int_0^{\infty}\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}. \end{aligned} \tag{4.19}$$

[43] Remember that iterating the EM algorithm gives a sequence of parameter estimates which converges to a (possibly local) maximum of the observed data log-likelihood (see Dempster et al. (1977)).

[44] As a consequence of the formula we state in eq. (B.16) in Appendix B.3, one can avoid the integration over the unbounded set $(0,\underline{x})\cup(\overline{x},\infty)$ by making the replacement

$$\big(1-p_{\mathrm{tr}}(\boldsymbol{\theta})\big)\int\limits_{(0,\underline{x})\cup(\overline{x},\infty)}\boldsymbol{s}(x)\frac{f_{\boldsymbol{\theta}}(x)}{1-p_{\mathrm{tr}}(\boldsymbol{\theta})}\mathrm{d}x = \frac{\partial \ln a(\boldsymbol{\theta})}{\partial\boldsymbol{\eta}} - \int_{\underline{x}}^{\overline{x}}\boldsymbol{s}(x)f_{\boldsymbol{\theta}}(x)\mathrm{d}x.$$

Proof of Theorem 1:

The first part follows from plugging the exponential family form $f_{\boldsymbol{\eta}}(x)$ into (4.17), and then deriving the first-order conditions with respect to $(\boldsymbol{\eta}', \nu')$. We carry out the details in Appendix B.3, where we also show that the first-order conditions are sufficient to find a maximum of $F(\boldsymbol{\eta}', \nu'|\boldsymbol{\eta}, \nu)$ in each iteration. The second part requires taking the second-order derivative of $L(\boldsymbol{\theta}, \nu)$ from (4.13) with respect to $\boldsymbol{\eta}$, and then rearranging with a formula based on Dempster et al. (1977, p. 5). This allows to express the derivative in terms of the expectation vectors and covariance matrices, see eqs. (B.16-B.17) in Appendix B.3. We also carry out the details of this derivation in Appendix B.3. Finally, the formulas in (4.18-4.19) are derived in Lemma 3 in Appendix B.1. □

Note that Theorem 1 provides the Hessian of the log-likelihood with respect to the natural parameters $\boldsymbol{\eta}$, not $\boldsymbol{\theta}$. Thus, the usual negative inverse of this matrix is the covariance matrix of $\hat{\boldsymbol{\eta}}$, not $\hat{\boldsymbol{\theta}}$. The delta method can be used to transform this into the covariance matrix of $\hat{\boldsymbol{\theta}}$. For this the Jacobian of the parameter transformation $\boldsymbol{\eta} \to \boldsymbol{\theta}$ is required, which we state for the lognormal distribution in Table B1 in Appendix B.3.

Finally, let us mention a consequence of the second equation for ν in part (1) of the theorem. Denoting with $(\hat{\boldsymbol{\theta}}, \hat{\nu})$ the maximum likelihood estimate, convergence of the EM iterations implies that $\hat{\nu} - k \approx \mathbb{E}_{(\hat{\boldsymbol{\theta}}, \hat{\nu})}[N - k|\mathcal{F}]$. By the law of iterated expectation and the properties of the Poisson distribution, this also implies that $\hat{\nu} - k \approx \mathbb{E}_{(\hat{\boldsymbol{\theta}}, \hat{\nu})}[\hat{\nu} U_k|\mathcal{F}]$. From this, a possible approximation to the mixed Poisson posterior of $N - k$ would be to ignore the randomness of $\hat{\nu} U_k$ and replace it by its expected value. This would give the approximation

$$U_k \approx \frac{\hat{\nu} - k}{\hat{\nu}} \Rightarrow (N - k)|\mathcal{F} \overset{\text{approx.}}{\sim} \text{Pois}(\hat{\nu} - k), \tag{4.20}$$

which illustrates how the Poisson distribution "acts like" a conjugate distribution for the total number of fields parameter in the likelihood function of the size-biased sampling model.

4.3.3 Models for the Discovery Times and Exploration Success

In this and the next section, we are not concerned with parameter estimation of $(\boldsymbol{\theta}, \nu)$, and hence drop the parameter subscripts and also drop the truncation factor by writing f for the truncated pdf f/p_{tr}.

In this section, we propose two approaches for integrating the size-biased sampling model with a stochastic model for the discovery times. The two approaches are conceptually somewhat different. The first approach (assumption 4a) derives the discovery times from directly extrapolating a (possibly nonlinear) trend function for the cumulative number of discovered fields, as a temporal function, into the future. The second approach (assumption 4b) is based on work from Herbert (1983) and Jakobsson et al. (2012) who have extended the size-biased sampling model into a model for the discovery yield per exploration effort. This approach also requires a trend extrapolation, now for the cumulative number of exploration wells, since a mapping of exploration wells onto time is required to derive the discovery times.

The next paragraph summarizes the essence of the extension from Herbert (1983) and Jakobsson et al. (2012).

Previous extensions to model the discovery yield per exploration effort

It is possible to extend the size-biased sampling model to include the unsuccessful exploration wells ("dry wells") that are drilled before a new oil fields is discovered. Formally, the extension from Herbert (1983) and Jakobsson et al. (2012) requires a new parameter, say b_0, which heuristically stands for the "empty area" in the region, i.e. the area that is explored for oil but does not contain oil fields. Let the variable $Z_i - Z_{i-1}$ represent the number of exploration wells that are drilled between the $(i-1)^{\text{th}}$ and the i^{th} discovery (so that there are $Z_i - Z_{i-1} - 1$ dry wells). Defining $b_i = x_i^\beta + \ldots + x_n^\beta$ when there are n fields in the region, the ratio $b_i/(b_i + b_0)$ can be interpreted as the probability that an exploration well strikes the "area" occupied by the remaining fields, b_i, rather than the "empty area", b_0. Extending (4.7) in this way, Herbert (1983, p. 36) arrives at an equation similar to

$$\mathbb{P}[X_1 \in \mathrm{d}x_1, \ldots, X_n \in \mathrm{d}x_n, Z_1 = z_1, \ldots, Z_n = z_n] = n! \prod_{i=1}^{n} f(x_i) \frac{x_i^\beta}{b_i + b_0} \Big(\frac{b_0}{b_i + b_0} \Big)^{(z_i - z_{i-1} - 1)} \mathrm{d}x_i.$$

In their Monte Carlo simulation, Jakobsson et al. (2012, p. 867) add "empty area" that is proportional to the simulated value of the total field size, so that the fields always cover 10% of the "total area". In our notation, this corresponds to setting $\beta = 1$ and $b_0 = 9b_1 = 9(x_1 + \ldots + x_n)$.

Modeling the discovery times via a pure birth process

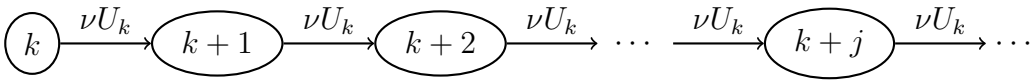
This subsection proceeds as follows. We first explain a few properties that the stochastic process needs to have in order to be a reasonable model for the discovery times. We then explain the two assumptions 4a and 4b, first in simple words, and afterwards fill in the formal details. For each assumption the explanations are organized as follows: first, we state the conditions under which it is consistent with assumption 3 and the resulting posterior for N, second, we relate the discovery times to random variables with fully specified distributions, and third, we propose our estimation strategy for the involved parameters.

In probability theory, a *counting process* is a plain type of stochastic process that only takes on nonnegative integer values and is nondecreasing over time, so that the increments of the process measure how many "events" occur in a given time interval. We let $(N_t)_{t \geq \underline{t}}$ denote the counting process for the number of discovered oil fields, so that the increment $N_t - N_s$ measures how many fields are discovered in the time interval $(s, t]$. Thus, the successive jump points at which $(N_t)_{t \geq \underline{t}}$ increases correspond to the successive discovery times. Because there are already $k \geq 0$ discoveries during the sample period $(0, \underline{t}]$ (see Figure 4.2), the process needs to start at $N_{\underline{t}} = k$.

In the previous section with assumption 3, we only had the variable N representing the total number of fields, while now with assumptions 4a or 4b we want to extend N into a stochastic process $(N_t)_{t \geq \underline{t}}$ that also lends a distribution to the discovery times. Thus, we assume that all fields in the region will be found eventually, which leads us to *redefine* $N \coloneqq \lim_{t \to \infty} N_t$. We therefore need to ensure that N *defined in this way* is consistent with assumption 3 and the resulting posterior as derived in sect. 4.3.2.

To create a counting process with enough modeling flexibility, we combine a textbook-type *pure birth process,* denoted as $(M_u)_{u \geq 0}$, with a function $\Lambda(t)$ that we call a "trend

Figure 4.3: Transition-rate diagram for assumption 4a



Note: The circles represent the states of the process (the number of discovered fields), the arrows show the transition / discovery rates.

function". Turning first to the trend function, we assume that it is defined on the domain $[\underline{t}, \infty)$, that it starts at $\Lambda(\underline{t}) = 0$, and that it is increasing and differentiable. We denote its inverse and derivative by Λ^{-1} and Λ', respectively. We combine this function with the pure birth process by making the *composition* $N_t \coloneqq M_{\Lambda(t)}$. The role of $\Lambda(t)$ is to introduce time-dependent behavior into $(N_t)_{t \geq \underline{t}}$.[45]

Turning now to the *pure birth process* $(M_u)_{u \geq 0}$, this process has the defining property that it is a continuous-time Markov chain which can only jump upward by one step at a time. It can be seen as an immediate generalization of the Poisson process. The process is fully characterized by a starting point $M_0 = k$ and a sequence of nonnegative numbers $R_{k+1}, R_{k+2}, \ldots$ which determines the transition probabilities over any small interval $(u, u+h]$ as $h \downarrow 0$. This is formally expressed by the following set of equations for each $i \geq k+1$ (see, e.g., Karlin and Taylor (1975), p. 119):

$$\begin{aligned} &\mathbb{P}[M_{u+h} - M_u = 1 | M_u = i-1] = R_i h + \mathrm{o}(h), \\ &\mathbb{P}[M_{u+h} - M_u = 0 | M_u = i-1] = 1 - R_i h + \mathrm{o}(h), \\ &\mathbb{P}[M_{u+h} - M_u > 1 | M_u = i-1] = \mathrm{o}(h). \end{aligned} \tag{4.21}$$

Here, $\mathrm{o}(h)$ is the usual "little-o" notation for terms which decline in h faster than linearly. The numbers $R_{k+1}, R_{k+2}, \ldots$ are called the *transition rates*, and as we have defined it, R_i is the rate for transitioning *away from state* $i-1$. Thus, R_i is the rate for discovering the i^{th} field, so we also use the term *discovery rate* for R_i in this section.[46]

Assumptions 4a and 4b now give a concrete form to the discovery rates and to the trend function.

Assumption 4a. *The discovery rates R_i in (4.21) are constant and independent of the field sizes that are yet-to-be discovered. Time-varying behavior of the process $(N_t)_{t \geq \underline{t}}$ is introduced (to allow for time-inhomogeneity in the distribution of the discovery times) and fully governed by the trend function $\Lambda_{\mathrm{a}}(t)$, which is required to be a bounded function.*

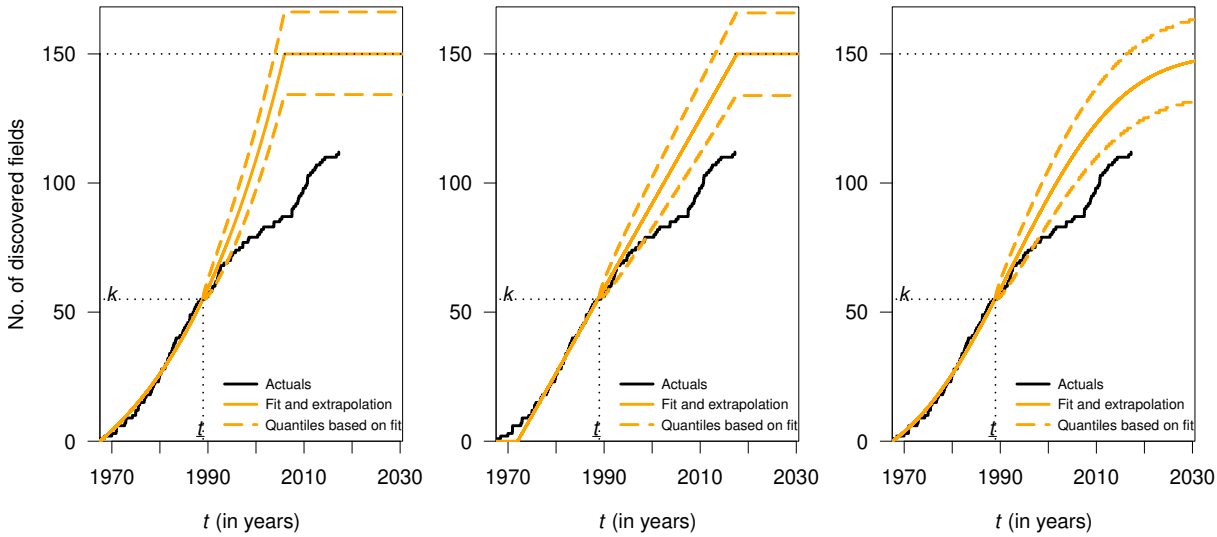
When transition rates are constant in (4.21), the pure birth process M_u becomes a Poisson process with intensity parameter equal to the constant transition rate (e.g., Karlin and Taylor (1975), p. 118; Shortle et al. (2018), p. 68). In order make assumption 4a

[45] In technical terms, $(N_t)_{t \geq \underline{t}}$ is said to be a (time-)inhomogeneous pure birth process or a pure birth process with nonstationary increments. Making a Taylor series expansion, the equations which correspond to (4.21) for N_t are

$$\begin{aligned} &\mathbb{P}[N_{t+h} - N_t = 1 | N_t = i-1] = R_i \Lambda'(t) h + \mathrm{o}(h), \\ &\mathbb{P}[N_{t+h} - N_t = 0 | N_t = i-1] = 1 - R_i \Lambda'(t) h + \mathrm{o}(h), \\ &\mathbb{P}[N_{t+h} - N_t > 1 | N_t = i-1] = \mathrm{o}(h). \end{aligned}$$

[46] We also use the term "discovery rate" in sect. 4.3.4 and sect. 4.5 to refer to the aggregate rate of discovered oil amounts in the region (e.g. measured in Mega-barrels per year). There should be no confusion, as we use the term "discovery rates" for the abstract transition rates R_i only in this section.

Figure 4.4: Illustration of the estimation strategy for assumption 4a



Note: The actuals (in black) represent Norwegian oil field discoveries, see section 4.4 for the data source. The trend functions that are fitted and extrapolated are: an exponential function (left), a linear function (middle), a cumulative Gaussian function (right). The sample period to which the functions are fitted covers the years 1967 until $\underline{t} = 1989$. As the value for the asymptote (the expected number of oil fields) we assume 150. 5%- and 95% quantiles are shown, which are computed using the approximation given in (4.20) with $U_k = (150 - k)/150$. $k = 55$.

consistent with assumption 3 and the resulting posterior for N, we need the following condition on the constant transition rate and the function $\Lambda_a(t)$.

Proposition 4a. *For* $N := \lim_{t\to\infty} N_t$*, it holds that* $(N - k)|U_k \sim \text{Pois}(\nu U_k)$ *if the product of* $\lim_{t\to\infty} \Lambda_a(t)$ *and the constant transition rate is equal to* νU_k*.*

A proof of Proposition 4a can be found in Appendix B.4.

Thus, we can further specify assumption 4a by setting $\lim_{t\to\infty} \Lambda_a(t) = 1$, and $R_i = \nu U_k$ for all $i \geq k + 1$, which follows from Proposition 4a without loss of generality.[47] Figure 4.3 shows the corresponding transition-rate diagram.

Turning to the distribution of the discovery times, first note that a homogeneous Poisson process like $(M_u)_{u\geq 0}$ can be characterized informally by saying that "it distributes points 'at random' over the infinite interval $[0, \infty)$ in much the same way that the uniform distribution distributes points over a finite interval" (Karlin and Taylor (1975), p. 123). Formally, this refers to the well-known order statistic result (Karlin and Taylor (1975), pp. 126f.; Feigin (1979); Puri (1982)). The case of $N_t = M_{\Lambda_a(t)}$ where $\Lambda_a(t)$ is a *bounded function* is a special case of this result. Here N_t simply distributes a finite number of points as the *order statistics* from $\Lambda_a(t)/(\lim_{t\to\infty} \Lambda_a(t))$. Thus, the following is a recipe for simulation: make $N - k$ ordered draws from the uniform distribution on $[0, 1]$, denoted by $\mathcal{U}_{k+1} < \ldots < \mathcal{U}_N$, and then compute the jump points of N_t by $T_i = \Lambda_a^{-1}(\mathcal{U}_i)$.

Since assumption 4a reduces to saying that, for given N, the distribution of the discovery times is fully governed by how the function $\Lambda_a(t)$ looks like, how can a parametric estimate

[47] Alternatively, one can set $\lim_{t\to\infty} \Lambda_a(t) = \nu U_k$ and $R_i = 1$.

of the function be derived from the previous discovery times? This can be done by fitting a parametric function to the observed step-function for the cumulative number of discovered fields, and extrapolating beyond $\underline{t}$ towards an asymptote. A reasonable estimate for this asymptote is the expected number of fields, $\mathbb{E}[N|\mathcal{F}]$. Letting $n(t)$ denote such a parametric function that is fitted to the data on $(0, \underline{t}]$ and extrapolated, we get the estimate of $\Lambda_{\mathrm{a}}(t)$ by $\Lambda_{\mathrm{a}}(t) = (n(t) - n(\underline{t}))/(\lim_{t\to\infty} n(t) - n(\underline{t}))$ for $t \in [\underline{t}, \infty)$, since this function has the limit one (for $t \to \infty$).

Figure 4.4 shows an example of the estimation strategy using data from Norwegian oil field discoveries. In the left panel $n(t)$ is chosen as an exponential trend function, in the middle as a (piecewise) linear trend function, and on the right as a cumulative Gaussian trend function. The sample period to which the functions are fitted covers the years 1967 until $\underline{t} = 1989$. As the value for the asymptote (the expected number of oil fields) we assume 150. For the exponential fit the rate at which new oil fields are discovered grows by 4% p.a., while for the linear fit on average 3.3 new fields are discovered p.a.. The Gaussian extrapolation is similar to the linear initially, but then negative feedback from approaching the asymptote slows down the pace of new discoveries. Figure 4.4 shows that all three extrapolations overpredict the actual pace at which new oil fields were found in Norway. The data indicate a structural break with a declining pace of discovery that started before 2000 and ended somewhat before 2010. The right panel of Figure C3 in Appendix C shows that exploration well drilling can at least partially explain this: in the midst of the decade 2000-2010 there was also a slow-down in the pace of exploration well drilling.

The example illustrates the difficulty of deciding on a "suitable" functional form for $n(t)$ / $\Lambda_{\mathrm{a}}(t)$, so this aspect bears considerable specification uncertainty and is not without bias, as Figure 4.4 shows. Nonetheless, a set of different functions can be interpreted as representing different scenarios for the pace of new discoveries. To improve the model, one could specify $n(t)$ / $\Lambda_{\mathrm{a}}(t)$ as a function that depends on an index of time-dependent explanatory variables instead of on time only, using variables that presumably influence the chance of discovery, such as the number of drilled of exploration wells. However, we do not pursue this approach of introducing exploration well drilling into assumption 4a in this paper. Instead we use assumption 4b, where we introduce the rate of exploration well drilling differently. Assumption 4b is still based on (4.21) and $N_t \coloneqq M_{\Lambda(t)}$, but now the discovery rates and the function $\Lambda(t)$ are specified differently.

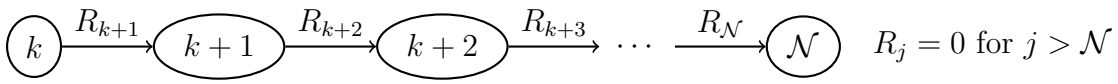
Assumption 4b. *At any stage in the discovery process, the discovery rates R_i in (4.21) are a function of the field sizes that are yet-to-be discovered. Additional time-varying behavior of the process $(N_t)_{t\geq\underline{t}}$ can occur due to temporal variation in exploration well drilling. For this, it is assumed that the cumulative number of drilled exploration wells follows the trend function $\Lambda_{\mathrm{b}}(t)$.*[48]

As Herbert (1983) and Jakobsson et al. (2012), we use functions of the yet-to-be discovered field sizes to model the probability with which a marginal unit of exploration effort discovers one of the remaining fields. We assume that these discovery rates are

$$R_i \coloneqq \begin{cases} \frac{X_i^\beta + \ldots + X_{\mathcal{N}}^\beta}{a_0(X_i^\beta + \ldots + X_{\mathcal{N}}^\beta) + b_0}, & \text{if } i = k+1, \ldots, \mathcal{N}, \\ 0, & \text{if } i \geq \mathcal{N}+1, \end{cases} \tag{4.22}$$

[48] As a generalization, $\Lambda_{\mathrm{b}}(t)$ could be an index of time-dependent explanatory variables including but not restricted to exploration well drilling.

Figure 4.5: Transition-rate diagram for assumption 4b



Note: The circles represent the states of the process (the number of discovered fields), the arrows show the transition / discovery rates.

where also

$$(\mathcal{N} - k)|U_k \sim \text{Pois}(\nu U_k).$$

The parameters a_0, b_0 are nonnegative and at least one of the two must be strictly positive. As above, the parameter b_0 represents "empty area" which does not diminish with exploration, while the new parameter a_0 also allows "empty area" to diminish with each discovery in proportion to the yet-to-be discovered field sizes. Figure 4.5 shows the transition-rate diagram for assumption 4b.

Next follows a technical condition that makes the assumption consistent with assumption 3 and the resulting posterior for N.

Proposition 4b. *If $\Lambda_\text{b}(t)$ is an unbounded function, the distribution of N_t will approach that of $\mathcal{N}$ for $t \to \infty$. In this case it holds for $N \coloneqq \lim_{t\to\infty} N_t$ that $(N - k)|U_k \sim \text{Pois}(\nu U_k)$.*

A proof of Proposition 4b can be found in Appendix B.4.

To distinguish more clearly between assumptions 4a and 4b, we now use the index z instead of u, i.e. M_u becomes M_z. Also, let the jump points of M_z be denoted by $Z_k, Z_{k+1}, \ldots, Z_N$, where by assumption $Z_k = 0$. A consequence of the definition in (4.21) is the property that the differences between the jump points are independent and exponentially distributed (e.g., Karlin and Taylor (1975), p. 121; Shortle et al. (2018), pp. 62ff.). We express this formally as $(Z_i - Z_{i-1}) \overset{\text{ind}}{\sim} \text{Exp}(R_i)$ for $i = k+1, \ldots, N$. A simulation thus requires only $N - k$ draws from the exponential distribution. From the relation between N_t and M_z, the jump points of N_t are then computed by $T_i = \Lambda_\text{b}^{-1}(Z_i)$. As a formal side note, note that the marginal distribution of Z_i is the general gamma distribution with parameters $R_{k+1}, \ldots, R_i$.

Turning now to our estimation strategy for a_0 and b_0, we first define $e_{k+1} \coloneqq \mathbb{E}\big[X_{k+1}^\beta + \ldots + X_N^\beta \big| \mathcal{F}\big]$, and $e_i \coloneqq x_i^\beta + \ldots + x_k^\beta + e_{k+1}$ for each $i = 1, \ldots, k$. For the first k discovered fields, we use data on the number of exploration wells that had to be drilled to discover a field around the time of their discoveries. Let this variable be denoted by $r_{\text{obs},i}^{-1}$, so that $r_{\text{obs},i}$ is the observed success rate of an exploration well around the time when the i^th field was discovered. Extending (4.22) backwards to $i = 1, \ldots, k$ and replacing $X_i^\beta + \ldots + X_N^\beta$ with its expected value given the data, e_i, we can rearrange the resulting equation into the linear regression specification

$$r_{\text{obs},i}^{-1} = a_0 + b_0 e_i^{-1} + \upsilon_i, \quad i = 1, \ldots, k, \tag{4.23}$$

where υ_i is an error term, and from which estimates $\hat{a}_0$ and $\hat{b}_0$ are readily computed by OLS.

For the function $\Lambda_\text{b}(t)$, we choose the parametric function $\Lambda_\text{b}(t) = ct^d - c\underline{t}^d$. The involved parameters can be estimated by least squares fitting of ct^d to the time series of the cumulative number of exploration wells drilled during $(0, \underline{t}]$.

4.3.4 Combination, Distributional Properties, and an Asymptotic Approximation

The first part of this section leads to Theorem 2, in which we summarize what the assumptions from sections 4.3.1-4.3.3 together imply for the distributional properties of the stochastic process $(Q_\tau)_{\tau \geq \underline{t}}$, presented in a way suitable for a Monte Carlo (MC) simulation. It is here where we introduce an alternative representation of the distribution of the size-biased variables $(X_1, \ldots, X_N)$, called by Pitman and Tran (2015) the distribution of *induced order statistics*. To let this paper be self-contained we fully derive this representation in Appendix B.5. There, we also show that the representation requires only a slight modification after conditioning on the first k variables $(X_1, \ldots, X_k)$ and Γ_k, which has not been shown in the reference Pitman and Tran (2015).

In the second part of this section, we explain and make use of an asymptotic result which Pitman and Tran (2015) derive from the induced order statistics representation. The result is a coupled system of *first-order differential equations,* which for large n (or ν) can be used to approximate the mean function of the regional discovery rate. We show how this result can be extended to the mean function of the production rate for a production profile like (4.2). We apply this result first to the case of assumption 4a, and then show how analogous equations can be derived for the case of assumption 4b. In the latter case, the model actually becomes close to the Arps-Roberts model, a connection which has already been recognized by Herbert (1983).

We conclude this section with a subsection where we verbally explain what kind of dynamics we expect for the overall model of production and discovery.

Distributional properties

To conduct a MC simulation of the remaining field sizes, one can start with drawing from the joint distribution defined in (4.14). A draw for $(\tilde{X}_{k+1}, \ldots, \tilde{X}_N)$ represents the sizes of the remaining fields, however arranged in an uninformative order instead of in the order of discovery. Then, the order of discovery can be obtained by iteratively drawing indices from the discrete probability distribution defined in (4.5).

Alternatively, the distribution that arises from size-biased sampling can be related to the *order statistics from the Laplace transform* defined in (4.11). This representation has been derived first, for the marginal distributions of the sequence $(X_i,\ i = 1, \ldots, n)$, by Barouch and Kaufman (1976a) (assertion 4 on p. 15 with a proof on pp. 60f.). The authors have used this to analyze the sequence of means under size-biased sampling, the "discovery decline curve" (Barouch and Kaufman (1976b); Kaufman (1993), p. 190). For a deterministic vector of sizes, it has been shown by Gordon (1983) that size-biased sampling is closely related to the order statistics of independent exponentially distributed rvs. For the case as in assumptions 1 - 2, where the size-biased sampling permutes an iid random vector, Pitman and Tran (2015) use the result by Gordon (1983) to derive the corresponding Laplace transform based expressions, as derived by Barouch and Kaufman (1976a) for the marginals, for the joint distribution of $(X_1, \ldots, X_n)$.

Pitman and Tran (2015) refer to the resulting distribution as that of *induced order statistics*. The first studies of induced order statistics are i.a. Bhattacharya (1974) and David and Galambos (1974). $(X_1, \ldots, X_n)$ are called the induced order statistics of some iid

sequence of bivariate random vectors, $\big((\tilde{X}_i, \tilde{\Gamma}_i),\ i = 1, \ldots, n\big)$, if they have the distribution of the $\tilde{X}$-component after reordering the sequence by the $\tilde{\Gamma}$-component (Pitman and Tran (2015), p. 2489). In other words, let $\Gamma_1 < \ldots < \Gamma_n$ be the rearrangement of $(\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_n)$ in ascending order. Then, the X_i that was associated with Γ_i in the original sequence of bivariate random vectors is the i^{th} induced order statistic. For our case of size-biased sampling applied to oil fields, this is the i^{th} field size. In bullet points *e)-f)* of Theorem 2, we state the exact distributions of Γ_i and $X_i|\Gamma_i$ which follow from the *induced order statistics result for size-biased sampling*. Note that in place of bullet points *e)-f)*, we could have characterized the distribution of the field sizes equivalently as explained in the first paragraph of this subsection.

Theorem 2 now characterizes the stochastic process $(Q_\tau)_{\tau \geq \underline{t}}$ by summarizing the distributional consequences from all assumptions made in sections 4.3.1-4.3.3.

Theorem 2.
Under assumptions 1 - 3 and 4a or 4b, the distribution of $(Q_\tau)_{\tau \geq \underline{t}}$, given the data $\mathcal{F} := \big\{\boldsymbol{X} = (x_1, \ldots, x_k), N \geq k\big\}$ and the discovery times $(t_1, \ldots, t_k)$, can be characterized as follows:

a) Eq. (4.1) becomes $Q_\tau = \sum_{i=1}^{k} q(\tau, t_i, x_i) + \sum_{i=k+1}^{N} q(\tau, T_i, X_i)$, where q is for example (4.2), (4.3), or (4.4).

b) To take into account the uncertainty inherent in the parameters $\boldsymbol{\theta}$ which are estimated from a sample, we add a distribution for $\boldsymbol{\theta}$. Based on the ML-estimation, we assume that $\boldsymbol{\theta}$ has the multivariate normal distribution centered at the ML-estimates $\hat{\boldsymbol{\theta}}$ and with the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$.[49]

c) $(Q_\tau)_{\tau \geq \underline{t}}$ depends on rvs which do not explicitly appear in the formula, namely Γ_k, and $\Gamma_{k+1}, \ldots, \Gamma_N$.

d) The mixing rv Γ_k has the data-dependent pdf $g^{(0)}_{\boldsymbol{\theta}}(\gamma|\mathcal{F}) \propto \mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)} g_{\boldsymbol{b}}(\gamma)$, $U_k := \phi_{\boldsymbol{\theta}}(\Gamma_k)$, and the number of undiscovered fields $N - k$ has the mixed Poisson posterior $(N-k)|U_k \sim \mathrm{Pois}(\nu U_k)$.

e) The rvs $(\Gamma_{k+1}, \ldots, \Gamma_N)$ are distributed as $N - k$ order statistics from the cdf $1 - \phi_{\boldsymbol{\theta}}(\gamma)/\phi_{\boldsymbol{\theta}}(\Gamma_k)$ supported on $[\Gamma_k, \infty)$.

f) For $i = k+1, \ldots, N$, the field sizes $X_i|\Gamma_i$ are independently distributed with pdf $h_{\boldsymbol{\theta}}(x_i|\gamma_i) \propto x_i^{\beta} \mathrm{e}^{-\gamma_i x_i^{\beta}} f_{\boldsymbol{\theta}}(x_i)$ on support $[\underline{x}, \overline{x}]$.[50]

g) The discovery times $(T_{k+1}, \ldots, T_N)$ are distributed as specified in assumption 4a or assumption 4b from sect. 4.3.3. In case of assumption 4b, this distribution depends on the yet-to-be discovered field sizes $(X_{k+1}, \ldots, X_N)$.

[49] This can be justified as follows: without having a prior distribution, and without being able to work out a posterior from an uninformative prior and the likelihood (4.13), the only information about such a posterior for $\boldsymbol{\theta}$ are the ML-estimates and the estimated covariance matrix.

[50] Monte Carlo draws from this pdf can be obtained by a simple accept/reject procedure based on drawing from the (truncated) Weibull distribution, and an accept criterion based on the function $x f_{\boldsymbol{\theta}}(x)$. We explain this below eq. (B.28) in Appendix B.5.

Proof of Theorem 2:

Bullet points *e)-f)* follow from the induced order statistics result as outlined above and proven in Appendix B.5. Bullet points *d)* and *g)* have already been derived / explained in sects. 4.3.2 - 4.3.3. □

Asymptotic approximation

In this subsection, we use a result from Pitman and Tran (2015) that allows to express the "mean dynamics" of the size-biased sampling model as a coupled system of *first-order differential equations* with known solutions. This system asymptotically (for large n or ν) approximates the mean function of the aggregate discovery rate, and we show how this can be used to approximate the mean function of the aggregate production rate when individual fields have production profiles such as (4.2). This result can be useful to foster a better intuition about the dynamics of the model. Also, it could be used as an alternative way to estimate some of the parameters of the model, namely by fitting the resulting curves for discovery or production to the actuals. For making forecasts, however, we want to stress that a MC simulation of the full stochastic model is clearly preferable because this yields a predictive distribution, which can account for some of the inherent uncertainties in the underlying variables (i.e. field sizes, discovery times).

To derive the result, it helps to see that the ρ-density and the h-density (see bullet point f above in Theorem 2) are related in a specific way. Making the substitution $\gamma = \phi^{-1}(u)$ for both densities, where ϕ^{-1} is the inverse of ϕ, let us define the new densities (Pitman and Tran (2015), p. 2499)

$$\begin{aligned} \rho^*(x|u) &:= \frac{\mathrm{e}^{-\phi^{-1}(u)x^\beta} f(x)}{u}, \\ h^*(x|u) &:= \frac{x^\beta \mathrm{e}^{-\phi^{-1}(u)x^\beta} f(x)}{-\phi'(\phi^{-1}(u))}, \end{aligned} \tag{4.24}$$

where ϕ' denotes the derivative of ϕ. Since h^* is a pdf it is easy to see that the following relation holds

$$-\phi'(\phi^{-1}(u)) = \int_{\underline{x}}^{\overline{x}} x^\beta u \rho^*(x|u) \mathrm{d}x.$$

Using this relation after taking the derivative of $u\rho^*(x|u)$, Pitman and Tran (2015, pp. 2499f.) show in lemma 19 that $\rho^*(x|u)$ satisfies the following differential equation (the key mathematical property here is the rule for the derivative of the inverse)

$$\frac{\partial}{\partial u}\big[u\rho^*(x|u)\big] = h^*(x|u) = \frac{x^\beta u \rho^*(x|u)}{\int_{\underline{x}}^{\overline{x}} x^\beta u \rho^*(x|u) \mathrm{d}x}, \quad u \in (0,1], \tag{4.25}$$

subject to the boundary condition $\rho^*(x|1) = f(x)$.

The result can be interpreted as follows: for any distribution $f(x)$ and a positive size measure $w(x)$, the distribution $w(x)f(x)/\mathbb{E}[w(X)]$ is called its size-biased distribution, or alternatively, its weighted distribution with weight function $w(x)$ (e.g., Patil and Rao (1978); Vardi (1982)). Equation (4.25) thus reveals that for the family of density functions $\rho^*(\cdot|u)$ over $u \in (0,1]$, $h^*(\cdot|u)$ is the size-biased density of $\rho^*(\cdot|u)$ with size measure $w(x) = x^\beta$.

It is well-known that in the limiting case (for $n \to \infty$) of size-biased sampling from an iid sequence with distribution f, "finite elements" of the sequence (i.e., with indices i such that $i/n \to 0$) become iid and have the size-biased distribution of f (e.g., Patil and Rao (1978); Vardi (1982); Pitman and Tran (2015), p. 2500). Based on their result in lemma 19, Pitman and Tran (2015) show in corollary 21 and theorem 23 that a Glivenko-Cantelli-type theorem holds for the last u-fraction from a size-biased sequence. In particular, the result is that the empirical distribution function of the last $\lfloor nu \rfloor$ elements from a size-biased sequence of length n converges for $n \to \infty$ to the distribution with density $\rho^*(\cdot|u)$, for all $u \in (0, 1]$.

Based on this result and the interpretation of the differential equation (4.25) provided by Pitman and Tran (2015, p. 2501), we propose here how one can derive an asymptotic approximation to the mean discovery rate, and more generally to the mean function of $(Q_\tau)_{\tau \geq \underline{t}}$. First, we need to specify how to map the variable that represents the remaining fraction of fields, u, onto the time axis. Under assumption 4a this is achieved by equating the two cdfs, $1 - \phi(\gamma)/\phi(\Gamma_k) \stackrel{!}{=} \Lambda_{\mathrm{a}}(t)$, and substituting $u = \phi(\gamma)$ and $U_k = \phi(\Gamma_k)$, thus yielding $u(t) = U_k(1 - \Lambda_{\mathrm{a}}(t))$. Let us denote by $y(x, t)$, $t \geq \underline{t}$ the following family of real-valued (as opposed to integer-valued) frequency distributions:

$$\begin{aligned} y(x,t) &\coloneqq \nu \cdot u(t) \cdot \rho^*(x|u(t)) \\ &= \exp\big\{ -\phi^{-1}(U_k(1 - \Lambda_{\mathrm{a}}(t)))x^\beta \big\} \nu f(x), \quad x \in [\underline{x}, \overline{x}], t \in [\underline{t}, \infty). \end{aligned} \tag{4.26}$$

Differentiating with respect to time is simply an application of (4.25) and the chain rule, which gives

$$\frac{\partial}{\partial t} y(x,t) = -\nu U_k \Lambda'_{\mathrm{a}}(t) \frac{x^\beta y(x,t)}{\int_{\underline{x}}^{\overline{x}} x^\beta y(x,t) \mathrm{d}x}. \tag{4.27}$$

Here we may replace the term νU_k with its expectation, $\mathbb{E}[N - k|\mathcal{F}]$. Note that eqs. (4.25-4.27) have a simple successive sampling interpretation. In (4.27), the aggregate frequency of fields declines at the rate of $\nu U_k \Lambda'_{\mathrm{a}}(t)$, and each infinitesimal size class $[x, x + \mathrm{d}x)$ assumes a certain fraction of that decline. This fraction equals the size measure x^β times the frequency of fields in the size class at time t, $y(x, t)\mathrm{d}x$, divided by the integral over all sizes classes.

Evaluating the initial condition $y(x, \underline{t})$ by taking the expectation given the data, we arrive at

$$\begin{aligned} \mathbb{E}\big[\nu U_k \cdot \rho^*(x|U_k)\big|\mathcal{F}\big] &= \mathbb{E}\big[\nu U_k \cdot \rho\big(x\big|\phi^{-1}(\phi(\Gamma_k))\big)\big|\mathcal{F}\big] \\ &= \mathbb{E}[N - k|\mathcal{F}] \cdot \int_0^\infty \rho(x|\gamma) g^{(1)}(\gamma|\mathcal{F}) \mathrm{d}\gamma, \quad x \in [\underline{x}, \overline{x}], \end{aligned} \tag{4.28}$$

which follows from equation (B.7) derived in Appendix B.1.

Note that multiplying (4.28) by x and integrating over $x \in [\underline{x}, \overline{x}]$ gives precisely the posterior mean of the undiscovered resources, see equation (4.35). Letting $y(x, \underline{t})$ start at this frequency distribution, the function $\big(- x \frac{\partial}{\partial t} y(x,t) \mathrm{d}x \big)$ is the discovery rate at any $t \geq \underline{t}$ for just the size class $[x, x + \mathrm{d}x)$. Hence, the aggregate discovery rate in the region at time t equals

$$\int_{\underline{x}}^{\overline{x}} -x \frac{\partial}{\partial t} y(x,t) \mathrm{d}x. \tag{4.29}$$

How can the aggregate production rate as a function of time be derived from this? Remember that the stochastic process defined in (4.1) just aggregates over all fields the

contributions q these fields make at time τ to the regional aggregate. Assume for now that the parameters of the production profiles are identical for all field-size categories, unlike in Table 4.1, and treat n and (t_i, x_i), $i = 1, \ldots, n$ as known quantities. Then, for all q-functions introduced in sect. 4.3.1, the formula for Q_τ becomes

$$\textstyle\sum_{i=1}^{n} q(\tau - t_i, 1)x_i.$$

Note that this is formally just like a kernel estimator where the realizations t_i occur at the "frequencies" x_i, and the kernel function being applied is $q(\cdot, 1)$ (e.g., the rectangular kernel in (4.4) or the production profile in (4.2)). The continuous analogue of this discrete situation is the same kernel function applied to all infinitesimal intervals, $[t, t + \mathrm{d}t) \subset [\underline{t}, \infty)$, where each interval "occurs" at the frequency given by the continuous function (4.29) times $\mathrm{d}t$. This naturally leads to the production rate as the convolution integral between $q(\tau - t, 1)$ and (4.29). The approach of computing production from convolutions based on past discoveries has already been used by Pukite et al. (2019).

Turning back to allow for size-dependent production profiles as shown in Table 4.1, there are simply two size categories (below vs. above 500 Mb) for which the calculation needs to be carried out separately and added. Letting $q_{\mathrm{prod}}^{(>500)}$ denote the function (4.2) for the above-500-Mb category, the aggregate production rate from this category at time τ equals

$$\sum_{i \in \{\{1,\ldots,k\}:\, x_i > 500\}} q_{\mathrm{prod}}^{(>500)}(\tau - t_i, 1)x_i + \int_{\underline{t}}^{\tau} q_{\mathrm{prod}}^{(>500)}(\tau - t, 1)\Big(\int_{500}^{\overline{x}} -x\frac{\partial}{\partial t} y(x,t)\mathrm{d}x \Big)\mathrm{d}t. \tag{4.30}$$

This approach can easily be extended if the production profiles differ for more than two size categories.

Finally, we derive a set of equations similar to (4.26-4.27) that have an analogous successive sampling interpretation for the case of assumption 4b where "empty area" is included. For this, consider first the following "extension" of the Laplace transform from (4.11),

$$\varphi(\gamma) \coloneqq a_0 \int_{\underline{x}}^{\overline{x}} \mathrm{e}^{-\gamma x^\beta} \nu f(x)\mathrm{d}x - b_0\gamma, \tag{4.31}$$

where the parameters a_0 and b_0 are as in sect. 4.3.3.

Note that $\varphi(\gamma)$ is decreasing and has range $(-\infty, a_0\nu]$, while $\varphi(\Gamma_k) - \varphi(\gamma)$ on the restricted domain $[\Gamma_k, \infty)$ is increasing and has range $[0, \infty)$, which is the same range that $\Lambda_{\mathrm{b}}(t)$ has. Thus, equating $\varphi(\Gamma_k) - \varphi(\gamma) \overset{!}{=} \Lambda_{\mathrm{b}}(t)$, solving for γ, and plugging this into $\mathrm{e}^{-\gamma x^\beta}\nu f(x)$, we define the family of frequency distributions, $y(x,t)$, $t \geq \underline{t}$ by

$$y(x,t) \coloneqq \exp\big\{ -\varphi^{-1}(\varphi(\Gamma_k) - \Lambda_{\mathrm{b}}(t))x^\beta \big\}\nu f(x), \quad x \in [\underline{x}, \overline{x}], t \in [\underline{t}, \infty). \tag{4.32}$$

Using $\Lambda_{\mathrm{b}}(\underline{t}) = 0$ and taking the expectation given the data, the initial condition is still equal to (4.28). Moreover, taking the time derivative of $y(x,t)$ gives the following differential equation (the key mathematical property to derive this is again the rule for the derivative of the inverse)

$$\frac{\partial}{\partial t} y(x,t) = -\Lambda_{\mathrm{b}}'(t) \frac{x^\beta y(x,t)}{a_0 \int_{\underline{x}}^{\overline{x}} x^\beta y(x,t)\mathrm{d}x + b_0}. \tag{4.33}$$

This equation has a successive sampling interpretation analogous to (4.27), but now the denominator mimics (4.22). In line with assumption 4b, the function $\Lambda'_{\mathrm{b}}(t)$ should now be a model of the drilling rate of new exploration wells.

Finally, note that the model (4.33) nests the Arps-Roberts model by setting $a_0 = 0$, since this yields the differential equation[51]

$$\frac{\partial}{\partial t} y(x,t) = -\Lambda'_{\mathrm{b}}(t) \frac{x^\beta y(x,t)}{b_0}. \tag{4.34}$$

Note that the linkage between the differential equations is removed in (4.34). Even with $a_0 > 0$, the Arps-Roberts model can be a reasonable approximation at any point t' where the parameter b_0 is sufficiently larger than a_0 times the "area" occupied by the fields, i.e. where $b_0 \gg a_0 \int_{\underline{x}}^{\overline{x}} x^\beta y(x,t')\mathrm{d}x$. Neglecting the a_0-term thus does not numerically change equation (4.33) by much. Since $y(x,t)$ is a declining function of time the Arps-Roberts model is then also good approximation for any $t \geq t'$. Under the initial condition consistent with (4.28) the Arps-Roberts model (4.34) has the solution $y(x,t) = \mathrm{e}^{-(\Gamma_k + \Lambda_{\mathrm{b}}(t)/b_0)x^\beta} \nu f(x)$.

Expectations regarding the dynamics of the model

In this subsection, we summarize what we expect about the dynamics of the overall model. The juxtaposition of the size-biased sampling model and the birth process model determines the rate of aggregate discovery in the region. We refer to this quantity in the following as the discovery rate, which should not be confused with the transition rates that we have also called discovery rates in sect. 4.3.3. It is clear that $\beta > 0$ implies that the expected field sizes decline with the succession of fields. The higher the parameter β the more rapid is the decline and the smaller are the variances of the field sizes conditional on discovery number. The strength of this relation also depends on the field-size distribution, where in particular the decline becomes more rapid if the variance of the field-size distribution is larger (Lee (2008), pp. 48 ff.).

For the case of assumption 4a, it is clear that if $\beta > 0$, and if the number of discoveries per time unit $\Lambda'_{\mathrm{a}}(t)$ (times $\mathbb{E}[N - k|\mathcal{F}]$) is a declining function, then the discovery rate is also a declining function. However, the function $\Lambda'_{\mathrm{a}}(t)$ can be increasing for a substantial portion of the whole discovery cycle in the region, consistent with an oil industry that grows its exploration activity at a certain rate per year. $\Lambda'_{\mathrm{a}}(t)$ can also be increasing if the exploration success rate increases via the "information effect" (Reynolds (1999); Rehrl and Friedrich (2006)). Thus, if $\Lambda'_{\mathrm{a}}(t)$ increases fast enough over some period the discovery rate will rise over this period.

When assumption 4b is made, the functional form in (4.22) implies that the exploration success rate, i.e. the chance that a unit of exploration effort discovers a field, declines monotonously. A large parameter a_0 has the effect of slowing down this decline initially, consistent with the "information effect", but the exploration success rate remains a declining function in any case. Of course, if the rate of exploration drilling $\Lambda'_{\mathrm{b}}(t)$ grows fast

[51] The original Arps-Roberts model is formulated not as a function of time but of cumulative exploration wells, and it uses data on the average area occupied by fields in each size class (e.g., Herbert (1982), p. 296; Schuenemeyer and Drew (1983), p. 154). Our formulation in (4.34) is a variant where cumulative exploration wells are mapped onto the time axis by $\Lambda_{\mathrm{b}}(t)$, and also where the average area of fields in size class $[x, x + \mathrm{d}x)$ is given by $w(x) = x^\beta$.

enough as to compensate this, the number of discoveries per time unit will rise. In the case of $\beta \gg 0$, though, once the average size per discovered field declines fast enough (call this date the regional "peak oil of discovery"), the discovery rate will also decline. These dynamics can well result in a single-peaked profile that is hump- or bell-shaped.

Of major interest is now how the discovery rate translates into the production rate. We have formally expressed the production rate in (4.30) as the convolution between the discovery rate and field-level production profiles, computed separately for different field-size categories and then added. First, note that the convolution is mathematically a smoothing operation, rendering the production rate more smooth or stretched-out. Second, remember from Figure 4.1 that the production profiles of giant fields decline more slowly than the profiles of smaller fields. The effect of this has been examined by Jakobsson et al. (2014), who sketch in figure 4 an example of the aggregation of three declining functions which have different decline rates and which are also shifted in time, and the result resembles the decline side of a bell-shaped curve. In summary, we expect that the production rate "inherits" the discovery rate profile, and that a bell-shaped pattern is likely if there are also enough larger fields with comparatively lower decline rates.

One element that we have left out in the above discussion of how discovery translates into production is the duration between discovery and first oil production. In our production profile model explained in sect. 4.3.1 this is the parameter δ_0, which we assume to be a fixed value for all fields in the region. This and the production profile implies that the production rate lags behind the discovery rate with a lag length greater than δ_0.[52]

Of course, the fact that an oil field has been discovered neither implies that it will be produced, nor determines when it will be produced. The fixed duration assumption can thus be criticized as unrealistic. For giant fields that are located offshore, Höök et al. (2009b, p. 51) calculate that the median duration between discovery and first oil has increased between the 1960s and the 2000s from 3.5 to 7.5 years. The duration between discovery and first oil is possibly affected by economic factors (such as the oil price), and regulatory, environmental, and technological factors, as well as the experience gained from previous development projects, all of which are in principle dynamic factors. On the other hand, as already mentioned in sect. 4.3.1, there is some evidence that field size does not have a strong influence on the duration between discovery and first oil (Mihalyi (2021), p. 14). For future research, the constant duration assumption could be refined by replacing it with a hazard-rate based duration model in which a time-dependent index captures the effects of some of these variables, including field size if this variable has a significant influence. In this paper, however, we abstain from this, in order to remove the additional layer of complexity and specification choices that would follow from making this refinement.

4.4 Data Sources

In this section, we present the sources for the data sets that are used in this paper. We also touch on some potential data issues that are related to the fact that the field-size data are based on remaining reserve estimates. In particular, we discuss the distinction between backdated and current reserves.

[52] The lag is greater than δ_0 because the fields require some time to reach their production plateaus (see Figure 4.1).

We obtain field-level oil reserve data for the U.S. part of the Gulf of Mexico (GOM) from the U.S. Bureau of Ocean Energy Management (BOEM).[53] We download the most recent data, which are as of 31th December 2019. Burgess et al. (2021) is the official publication where the data are described.[54] The main Excel table is a data set of all oil and gas fields in the GOM region, which contains for each field an estimate of the size ("original reserves") in million barrels (Mb), the discovery year,[55] water depth in feet (ft), and a classification into oil and gas fields. We also download the "Access database of field monthly production", from which we compute yearly production time series, and download a well-level data file ("mv_boreholes"), which contains all exploration wells and the date when each well reached its total depth. Based on this we calculate the number of exploration wells drilled per year.

In our analysis we only include the fields that are classified as oil fields.[56] In principle, the analysis could have been conducted with all fields using the sum of oil and gas reserves expressed in oil equivalents, but we restrict it to oil fields to avoid potential bias from mixing oil and gas fields. Important for our empirical analysis is the insight that discovery and production in the GOM region is best treated as two separate cycles. Burgess et al. (2021) plot in figure 15 annual oil production for all fields located in deep water (defined as deeper than 1,000 ft) and those in shallow water as two separate curves. The curves show that the shallow water fields started to produce a couple of decades before the deep water fields. Accordingly, we assign the 288 oil fields for which we obtain data to the two sub-regions, which we refer to as "GOM Flat" and "GOM Deep". Of the 288 oil fields, 164 are assigned to GOM Flat and 124 to GOM Deep. All of our calculations in sect. 4.5 are carried out for GOM Flat and GOM Deep as separate regions.

The data for Norway are retrieved from the Fact Pages of the Norwegian Petroleum Directorate (NPD), see https://factpages.npd.no/en. We obtain the field-level data sets "Overview", which contains the date when the discovery wellbore was completed, "Reserves", which contains the NPD's estimates of the field sizes ("original recoverable oil") in million standard cubic meters (Sm3), "Description", which contains the water depth in meters, and "Production - Saleable - Monthly - by field", which contains the net amount of oil produced by field and month in million Sm3. We convert the size and oil production data from million Sm3 to Mb,[57] and also convert the water depth from meters to feet in order to make the units comparable with those used in the GOM data set. Our download for the production data covers all years up to and including 2021. Our download for the field-size estimates reflects the NPD's reserve estimates as of 31th December 2020. Finally, we download the table "Wellbore - Statistics - Completion year (include junked)", from which we obtain the number of exploration wells drilled in each year.[58]

Regarding the field-size data, which consist of remaining reserve estimates plus the cumulative historical production, it is important to distinguish the two concepts of current

[53]The data are available at: https://www.data.boem.gov/

[54]See https://www.boem.gov/oil-gas-energy/resource-evaluation/discovered-resources/

[55]Since the discovery dates for the GOM region are available only up to the year, we obtain ties in the discovery order. We solve these ties by randomly assigning the discovery order for all fields with the same discovery year.

[56]We also delete 5 oil fields where either no data on original reserves or production is available. The total resulting number of oil fields with data is 288.

[57]The conversion factor is 1 Sm3 = 6.2898 barrel.

[58]There are actually two exploration well series, wildcats and appraisal wells. We use the sum of both to measure exploration wells.

vs. backdated reserves. Either of the two concepts can be found in applications where regional oil reserves or discovery rates are calculated and plotted as functions of time. When the current reserve concept is used, any revision in the reserve estimate of a field is assigned to the year it is made, while when the backdated reserve concept is used, any revision in the reserve estimate of a field is assigned to the discovery date of that field (Sorrell and Speirs (2009), pp. 10f.). In this way, discovery rates computed using the current reserve concept will lag behind discovery rates computed using the backdated reserve concept. Equations (4.3-4.4) show what the backdated reserve concept means mathematically: there is only a single increase in the reserves which occurs at the time of discovery t and is of magnitude x, the estimated field size. Thus, we use the backdated reserve concept in this paper.[59]

Moreover, instead of the estimates for recoverable oil it may be favorable to use estimates for in-place oil which also include unrecoverable oil. This approach would then require a model that converts in-place field sizes to recoverable field sizes (e.g., by applying a typical recovery factor) before these are plugged into the production profile. Estimates for in-place oil are indeed available for Norway but not for the GOM region, so this approach is not pursued in this paper. The approach remains a fruitful avenue for future research since oil-in-place is, in contrast to reserves, a geological fact, and thus oil-in-place estimates might contain less reporting bias (see also Jakobsson et al. (2012), p. 862).

4.5 Results

This exposition of the results is structured as follows. In sect. 4.5.1 we first provide a descriptive analysis of the empirical field-size distributions, and then analyze the correlations between the variables discovery order, field size, and water depth. In sect. 4.5.2, we present our estimation results for the parameters of the size-biased sampling model for all three regions and using several different sample periods. We compute the resulting estimates of undiscovered resources and compare these to the estimates that are published by the respective official agencies (i.e. the BOEM and the NPD). Finally, sect. 4.5.3 presents our main results, the MC simulation results for the overall model. In particular, we show here graphically how the distributions of the following stochastic processes evolve over time: the number of discovered fields, the discovered amounts (i.e. the discovery rate), and the production rate.

4.5.1 Data Description and Pre-Analysis

This section starts with a descriptive analysis of the empirical field-size distribution for each region. The aim is to find out whether the lognormal or the Pareto distribution is the better candidate for this distribution. Afterwards we analyze the correlation between the discovery order and the two variables field size and water depth, in order to give a first quantification to which degree the phenomenon of size-biased sampling is present, and to which degree a sampling bias due to water depth is also present.

[59] It would be entirely possible to alter the q-functions from sect. 4.3.1 so that they become consistent with the repeated reserve revisions of the current reserve concept. For this, include into eqs. (4.3-4.4) a model how reserve estimates get revised over time (typically they grow over time). Also, if it is plausible that not only ultimate recovery but also reported current reserves determine how fast an oil field is produced, this can be included in the production profile (4.2).

Table 4.2: Selected years for the empirical analysis

GOM Flat			Norway			GOM Deep		
Selected year	No. of disc. fields	Disc. resources (in Gb)	Selected year	No. of disc. fields	Disc. resources (in Gb)	Selected year	No. of disc. fields	Disc. resources (in Gb)
1957	17	3.55	1979	23	14.39	1990	26	4.19
1962	28	4.62	1984	40	22.90	1995	37	5.24
1967	47	6.22	1989	55	27.32	2000	61	8.05
2020	164	10.34	2022	112	35.46	2020	124	12.75

First we show in Table 4.2 several selected years upon which our empirical analysis is based. The table reports the number of oil fields and the amounts of oil discovered before a certain year. In Appendix C, we also show graphically how the actuals of the following variables evolve over time: the yearly oil and gas production time series (Figure C1), the cumulative amounts of discovered oil, the cumulative number of discovered oil fields, and the cumulative number of exploration wells (Figures C2-C4).

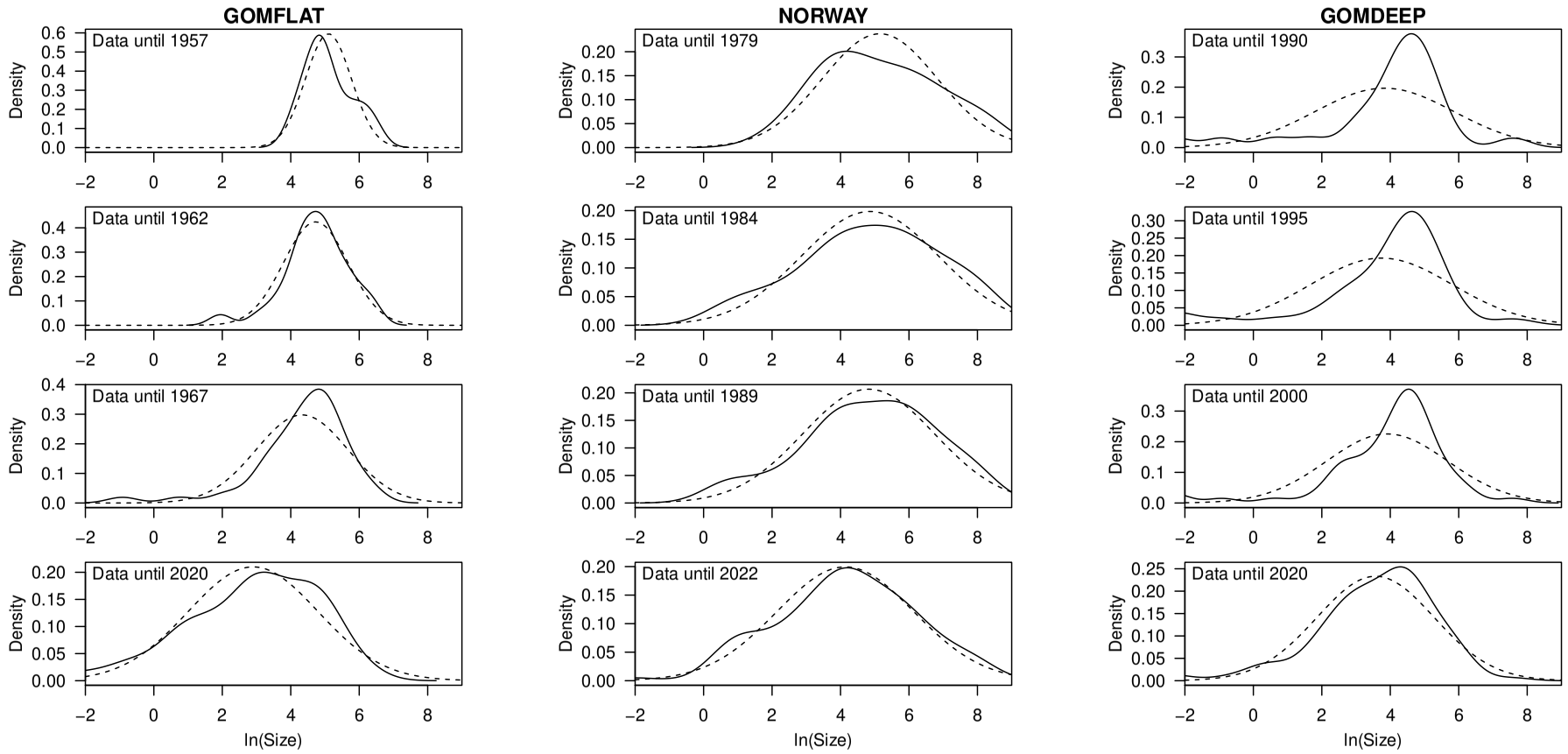
In contrast to the other two regions, we can assume for GOM Flat that in 2020 the discovery cycle for the region is completed, since the last new oil field is discovered in 2008. Overall, the total discovered oil in this region is quite small with only 10.34 Gb. However, this is also due to the fact that GOM Flat is predominantly a gas-prone region (BOEM (2017), p. 12). From Figure C1 one can see that gas production is substantially above oil production for GOM Flat, while for GOM Deep oil is dominant, and for Norway both is similar. Since oil and gas production pretty much move in lockstep for GOM Flat, the results we obtain for oil may be roughly transferable to gas, making our results for this region quantitatively more relevant.

Figure 4.6 depicts the empirical size distributions (with sizes in logarithms) for several subsamples of oil fields. A subsample for a given region contains all fields that are discovered before a selected year, see Table 4.2. To analyze the goodness-of-fit of the lognormal distribution, a fitted normal density is added in each panel as a dashed line. Also, we present descriptive statistics and test results for the null hypothesis of a lognormal distribution in Table 4.3.

For GOM Flat, what is clearly visible from Figure 4.6 is that the size distribution is sharply concentrated at larger values initially, while in later years it successively stretches out to smaller values. The distribution also shows a right tail that is above the normal fit for 1957. Together, this provides clear evidence of size-biased sampling. Figure 4.6 shows for 2020 that the normal distribution provides an acceptable fit, although the empirical size distribution has some degree of negative skew. Table 4.3 shows for 2020 that this deviation from the normal fit causes the AD- and KS-test to reject the null of a lognormal distribution at the 5% significance level.

Turning now to Norway, we see in Figure 4.6 that the size distribution is quite broad already in the earlier periods 1979-1989, substantially broader in comparison to the other regions. Also, the right tail of the size distribution is initially clearly above the normal fit. A considerable share of fields that are discovered before 1979-1989 have log-sizes above 6.2, which qualifies these fields as giant oil fields (since $\exp(6.2) \approx 500$ Mb). Then, between 1989 and 2022, most probability mass is added at or below the mean log-size

Figure 4.6: Plots of the ln(Size) distributions



Note: The solid line is a kernel density estimate, and the dashed line is a fitted normal density.

Table 4.3: Summary of the ln(Size) distributions

	Data Until	Min.	Max.	Mean	Sdev.	Skew.	Kurt.	AD (p-val.)	KS (p-val.)
GOM Flat	1957	4.13	6.39	5.10	0.67	0.53	2.28	0.24	0.17
	1962	1.95	6.39	4.73	0.94	-0.68	4.03	0.57	0.33
	1967	-0.92	6.39	4.33	1.34	-1.63	6.92	<0.01	<0.01
	2020	-2.30	6.39	2.90	1.90	-0.57	2.84	<0.01	0.02
	Data Until	Min.	Max.	Mean	Sdev.	Skew.	Kurt.	AD (p-val.)	KS (p-val.)
Norway	1979	2.12	8.21	5.15	1.68	0.24	2.12	0.70	0.58
	1984	0.76	8.21	4.86	2.01	-0.25	2.35	0.76	0.93
	1989	0.37	8.21	4.82	1.93	-0.35	2.59	0.60	0.76
	2022	-2.07	8.21	4.14	2.00	-0.23	2.84	0.50	0.66
	Data Until	Min.	Max.	Mean	Sdev.	Skew.	Kurt.	AD (p-val.)	KS (p-val.)
GOM Deep	1990	-2.30	7.59	3.83	2.03	-1.41	5.13	<0.01	<0.01
	1995	-2.30	7.59	3.74	2.07	-1.48	5.19	<0.01	<0.01
	2000	-2.30	7.59	3.91	1.77	-1.62	6.77	<0.01	<0.01
	2020	-2.30	7.59	3.57	1.71	-0.86	4.30	<0.01	0.17

Note: Kurt.=Kurtosis, AD=Anderson-Darling test, KS=Lilliefors (Kolmogorov-Smirnov) test.

of 4.14 ($\approx$ 65 Mb). This pattern that a large fraction of the giants are discovered early while mostly small- to medium-sized fields are discovered later provides clear evidence of size-biased sampling. Table 4.3 shows that the fit of the lognormal distribution is fairly good: the null of a lognormal distribution can neither be rejected by the AD-test nor by the KS-test in any of the years.

For GOM Deep, Figure 4.6 shows that the distribution is initially quite narrow and peaked, concentrating around a median log-size of about 4.4 ($\approx$ 82 Mb) for 1990-2000. Gradually it becomes stretched-out, also on the right tail, which implies that also several larger fields are discovered in later years. This pattern provides some evidence against size-biased sampling as the main mechanism that determines the discovery order. For 2020, a comparison of the empirical size distribution with the lognormal fit leads to similar conclusions as for GOM Flat, and Table 4.3 shows an inconclusive test result based on the AD- and the KS-test.

Overall, Figure 4.6 indicates that the lognormal distribution is a better candidate than the Pareto distribution for all three regions. Besides having more probability mass at small values, the Pareto distribution also has a thicker right tail compared to the lognormal distribution, which is quantitatively more relevant since this implies more very large fields (Jakobsson et al. (2014), p. 120). With the possible exception of Norway, a thicker right tail is not observable for the size distributions in the most recent periods shown in Figure 4.6.

In order to quantify the magnitude with which each of the two variables, size and water depth, is correlated with the discovery order, we report in Table 4.4 results from univariate regressions of the logarithm of each on the discovery rank. To disentangle the correlations between the two variables and the discovery order, we report in Table 4.5 the Spearman rank correlations between discovery date, size and water depth.

Table 4.4 shows that for the fields discovered before 2020 in GOM Flat, each successive field became on average 2.8% smaller. At the same time, the coefficient of ln(Depth)

Table 4.4: Results from regressing size and water depth on discovery rank

		Data Until 1957		Data Until 1962		Data Until 1967		Data Until 2020	
		ln(Size)	ln(Depth)	ln(Size)	ln(Depth)	ln(Size)	ln(Depth)	ln(Size)	ln(Depth)
GOM Flat	Rank	-0.033	0.043**	-0.058**	0.023**	-0.047***	0.027***	-0.028***	0.007***
		(0.037)	(0.019)	(0.028)	(0.010)	(0.014)	(0.005)	(0.002)	(0.002)
	Obs.	17		28		47		164	
		Data Until 1979		Data Until 1984		Data Until 1989		Data Until 2022	
		ln(Size)	ln(Depth)	ln(Size)	ln(Depth)	ln(Size)	ln(Depth)	ln(Size)	ln(Depth)
Norway	Rank	-0.003	0.032***	-0.047*	0.015**	-0.022	0.014***	-0.023***	0.008***
		(0.058)	(0.010)	(0.027)	(0.007)	(0.016)	(0.004)	(0.005)	(0.002)
	Obs.	23		40		55		112	
		Data Until 1990		Data Until 1995		Data Until 2000		Data Until 2020	
		ln(Size)	ln(Depth)	ln(Size)	ln(Depth)	ln(Size)	ln(Depth)	ln(Size)	ln(Depth)
GOM Deep	Rank	0.024	0.047***	0.006	0.025***	0.014	0.019***	-0.007*	0.008***
		(0.051)	(0.012)	(0.027)	(0.008)	(0.012)	(0.003)	(0.004)	(0.001)
	Obs.	26		37		61		124	

Note: Obs.=Number of observations. HAC (Newey/West) standard errors are in parenthesis. *p<0.1. **p<0.05. ***p<0.01.

Table 4.5: Rank correlations (Spearman) between size, water depth and discovery date

		Data Until 1957		Data Until 1962		Data Until 1967		Data Until 2020	
		Disc	Size	Disc	Size	Disc	Size	Disc	Size
GOM Flat	Size	-0.169		-0.421		-0.442		-0.718	
	Depth	0.493	-0.037	0.448	-0.150	0.586	-0.060	0.356	-0.143
		Data Until 1979		Data Until 1984		Data Until 1989		Data Until 2022	
		Disc	Size	Disc	Size	Disc	Size	Disc	Size
Norway	Size	0.005		-0.232		-0.157		-0.369	
	Depth	0.544	0.128	0.336	0.268	0.400	0.289	0.406	0.101
		Data Until 1990		Data Until 1995		Data Until 2000		Data Until 2020	
		Disc	Size	Disc	Size	Disc	Size	Disc	Size
GOM Deep	Size	0.050		0.005		0.043		-0.229	
	Depth	0.732	0.170	0.543	0.234	0.595	0.253	0.530	0.247

shows that each successive field was located on average in 0.7% deeper waters. These effects start to become statistically significant from the period 1962 onwards, and have initially even higher magnitudes. In Norway, the coefficients for the 2022 period are of similar magnitude: fields discovered before 2022 became 2.3% smaller and were located in 0.8% deeper waters, on average. Note that the ln(Size) coefficient for Norway is (marginally) significant before 2022 only in 1984, although in 1989 it has already the "correct" magnitude from 2022.

For GOM Deep, the coefficient of ln(Size) starts to become negative and (marginally) significant only in 2020, and has a small magnitude compared to coefficients from the other regions. In contrast, the coefficient of ln(Depth) is always significant and of roughly similar magnitude as in the other regions. When comparing the magnitudes of the ln(Depth)-coefficients across the regions, however, note that the water depth of the oil fields ranges from 1000-9000 ft in GOM Deep, while it ranges from 10-1000 ft in GOM Flat and from 200-1300 ft in Norway. Thus, the GOM Deep coefficients correspond to a much greater absolute increase in water depth.

Based on Tables 4.3-4.5 and Figure 4.6, the joint influence of size and water depth on the discovery order in GOM Deep may be summarized as follows. Primarily, fields in more shallow water are biased for early discovery. Secondarily, a size bias is observable when considering all data until 2020. Also, the rank correlation between size and water depth in Table 4.5 is throughout positive which reveals that larger fields tend to be located deeper. Thus, the initially "narrow" distribution observable in the right column of Figure 4.6 can be explained as follows: the bias to discover first fields in more shallow waters prevents the early discovery of several of the larger fields that are "hidden" in deeper waters. At the same time, the presence of at least some degree of size-biased sampling prevents also the early discovery of many of the smaller fields.

4.5.2 Estimation Results for the Size-Biased Sampling Model

In this section, we first explain some computational details regarding the parameter estimation. From this it will be clear which parameters are estimated rigorously from the data, and which parameters are "hyperparameters", by which we mean that these are fixed at certain values prior to the estimation (and thus are not rigorously estimated). Afterwards, we discuss our estimation results reported in Tables 4.6-4.7, in particular the implied undiscovered resources estimates for each region, which can be compared to point estimates published by the BOEM and the NPD.

Based on the analysis from the previous section we assume for each region that $f_{\boldsymbol{\theta}}/p_{\mathrm{tr}}(\boldsymbol{\theta})$ is the truncated lognormal distribution with parameters $\boldsymbol{\theta} = (\mu, \sigma)$ supported on $[\underline{x}, \overline{x}]$. We choose the support for each region on a rule-of-thumb basis from the observed field sizes. The values for the more important upper truncation point $\overline{x}$ are shown in Tables 4.6-4.7. For this distribution, we repeat the ML-estimation via the EM-algorithm as described in Theorem 1 for multiple grid points of the discoverability parameter β.[60] For each β, ML-estimates can be obtained by iterating the EM-algorithm from an initial guess (μ_0, σ_0, ν_0) until the sequence of log-likelihoods $L(\mu_0, \sigma_0, \nu_0)$, $L(\mu_1, \sigma_1, \nu_1)$, ... converges at some specified tolerance.

[60] We choose the grid points at a distance of 0.05, and adjust the range so that it accommodates a clear maximum of the log-likelihood function.

However, instead of the joint estimation of (μ, σ) and ν as stated in Theorem 1, we have decided to repeat the estimation with a few pre-determined values for ν, and then choose ν based on the likelihood.[61] Treating ν as a hyperparameter, our estimation procedure results in the conditional ML-estimates $\hat{\mu}(\beta)$ and $\hat{\sigma}(\beta)$ and their covariance matrix. To reduce the potential influence of the initial guess (μ_0, σ_0), we repeat the EM-algorithm for four different guesses and take the mean of the resulting estimates. After calculating the log-likelihood function $L(\beta)$ as in (4.13), the ML-estimate for β is easily obtained via one-dimensional grid search.

The integrals over x which appear in Theorem 1 are numerically calculated using the R-function "quadv" from the R-package "pracma", which implements a vectorized adaptive Simpson quadrature. Since the evaluation of the density $g_{\boldsymbol{b}}(\gamma)$ is costly, we evaluate all functions of γ on a pre-determined grid of values (see Appendix B.2 for a detailed description how we calculate $g_{\boldsymbol{b}}(\gamma)$). The integrals over γ are calculated numerically using these grid points, and we implement this via Simpson's 1/3 rule.

The conditional ML-estimates $\hat{\mu}(\beta)$, $\hat{\sigma}(\beta)$ allow to calculate the posterior mean of the undiscovered resources as a function of β by the formula

$$\mathbb{E}_\beta\Big[\sum_{i=k+1}^{N} X_i \Big| \mathcal{F}\Big] = \mathbb{E}_\beta[N-k|\mathcal{F}] \cdot \int_0^\infty \int_{\underline{x}}^{\overline{x}} x \cdot \rho_\beta(x|\gamma) g_\beta^{(1)}(\gamma|\mathcal{F}) \mathrm{d}x \mathrm{d}\gamma. \tag{4.35}$$

For several specifications for each region, we plot the conditional ML-estimates $\hat{\mu}(\beta)$, $\hat{\sigma}(\beta)$, $L(\beta)$, and the function (4.35) in Appendix C (see Figures C8-C11). The figures show that the mean of the undiscovered resources is a declining function of β at nearly all grid points, as expected because a higher β corresponds to a higher degree of size-bias in the successive sampling scheme. In particular, $\beta = 0$ is equivalent to random sampling for which the mean of the total resources should be close to $\frac{\nu}{k}\sum_{i=1}^{k} x_i$.

The monotony of the function (4.35) provides an alternative approach to choose the parameter β if a point estimate of the ultimately recoverable resources (URR) in the region is available from external sources. In this case, one can choose β so that the mean undiscovered resources plus the resources from known fields is closest to the external estimate. This is similar to what is known in the discovery process literature as the "anchored method", see for example Chen and Sinding-Larsen (1999, p. 50). In this terminology, we "anchor" the discoverability parameter β as to match the external URR point estimate.[62] We pursue this approach as a further variant in addition to the ML-estimation of β. For GOM Flat, we use the value of total discovered oil resources by 2020 (10.34 Gb). For Norway, we refer to the most recent estimate of total oil resources on the Norwegian continental shelf published in NPD (2021), which amounts to about 50 Gb.[63] For GOM Deep, we refer to the resource assessment from the U.S. BOEM

[61] For the joint estimation of (μ, σ) and ν we have obtained parameter estimates which are quite dependent on the initial guess and sometimes yielded nonsensical estimates (e.g. ν very low or high). To simplify the calculations we have thus decided to fix the values for ν. This should, however, not discourage future research to carry out the joint estimation. The above-mentioned problems could potentially be avoided by imposing lower and upper bounds on the parameters in each iteration of the EM algorithm.

[62] The match is not exact since (4.35) has to be computed numerically, for which choose β on a grid with distance of 0.05.

[63] The report is available for download at: https://www.npd.no/en/facts/publications/reports/resource-accounts/. The 50 Gb figure is obtained from the second column ("Oil million scm") of table 1-1. We deduct from the total the category "contingent resources in fields", as these should

Table 4.6: Estimation results

	GOM Flat, $\underline{t} = 1962$		GOM Flat, $\underline{t} = 1967$		Norway, $\underline{t} = 1984$		Norway, $\underline{t} = 1989$	
$\overline{x}$	600	600	600	600	4400	4400	4400	4400
ν	164	164	164	164	150	150	150	150
	ML-EST	FIT-URR	ML-EST	FIT-URR	ML-EST	FIT-URR	ML-EST	FIT-URR
$\hat{\beta}$	1.50	1.10	1.05	0.60	0.50	0.25	0.50	0.20
$\hat{\mu}$	2.96	3.40	2.28	3.30	1.65	3.64	2.47	4.11
	(0.17)	(0.24)	(0.26)	(0.29)	(0.72)	(0.33)	(0.30)	(0.32)
$\hat{\sigma}$	1.58	1.41	2.64	2.09	3.29	2.90	3.00	2.61
	(0.22)	(0.17)	(0.65)	(0.31)	(0.76)	(0.51)	(0.50)	(0.32)
Res[a]	3.52	5.63	1.51	4.43	8.11	25.21	5.87	23.84
	[2.67;4.44]	[3.81;7.56]	[0.98;2.14]	[2.85;6.18]	[4.30;12.71]	[13.19;37.95]	[3.19;9.22]	[13.45;35.15]
L[b]	-166.41	-167.36	-272.59	-276.24	-274.45	-274.92	-374.16	-375.02
LR[c]	1.90		7.30***		0.94		1.72	

Note: Estimated standard errors are in parenthesis.

[a]Undiscovered resources in Gb. The upper row is the mean estimate, in brackets are the 10% and 90% quantiles.

[b]Log-likelihood evaluated at the parameter estimates. [c]Likelihood ratio statistic. *p<0.1. **p<0.05. ***p<0.01.

published in BOEM (2017). Therein, estimates of undiscovered technically recoverable oil and gas resources in the Gulf of Mexico are presented per geological assessment unit (see table 11 of BOEM (2017) on p. 61). Choosing the assessment units labeled as "slope" which correspond to deep water, we arrive at a resulting mean estimate of undiscovered technically recoverable oil resources of about 40 Gb.[64] Adding this to the 13 Gb discovered until 2020 yields 53 Gb, a similar figure as for Norway. We label the specifications where we choose $\hat{\beta}$ as to best match the above-mentioned URR figures "FIT-URR", while when we choose $\hat{\beta}$ by maximizing the likelihood function we label this "ML-EST".

Tables 4.6-4.7 summarize our estimation results for all regions and all specifications. A specification consists of, (1) the year $\underline{t}$ that separates the sample period from the projection period, (2) whether $\hat{\beta}$ is determined by "ML-EST" or "FIT-URR", and (3) the hyperparameters ν and $\overline{x}$, which are, however, not varied in the specifications shown in Tables 4.6-4.7. The original plan was to pick as $\underline{t}$ for each region the three latest years that are shown in Table 4.2. However, for GOM Deep we have obtained estimates with huge variances for the year $\underline{t} = 1995$, so we start with $\underline{t} = 2000$ and add the year $\underline{t} = 2005$. For GOM Flat, we omit estimation for $\underline{t} = 2020$ because discovery in the region can be assumed as completed. Besides the parameter estimates, the tables show in the row "Res" our estimates of undiscovered resources with associated 10%- to 90%-quantiles in brackets.[65] The resource estimates and quantiles are computed from MC simulations of the undiscovered field sizes.

Tables 4.6-4.7 show that comparatively the highest estimates for β are obtained for

belong to the already discovered fields, and convert the resulting number from standard cubic meters to barrels.

[64]We include also the large "Norphlet" play which has many deep water oil fields (BOEM (2017), pp. 55f.), and exclude the "Mezozoic Slope" which is considered as "conceptual" (BOEM (2017), p. 56). Also, we deduct 5% from the reported figures to arrive at the 40 Gb figure. The reason is that we have deleted from our sample the fields which are classified as gas fields. We calculate from all fields discovered before 2020 in GOM Deep that 5% of oil resources are actually contained in the fields which are classified as gas fields.

[65]We use the 10%- and 90%-quantiles since these correspond to the established concepts of 1P (proved) and 3P (proved+probable+possible) estimates of reserves (e.g., Thompson et al., 2009, p. 12)

Table 4.7: Estimation results (cont'd)

	Norway, $\underline{t} = 2022$		GOM Deep, $\underline{t} = 2000$		GOM Deep, $\underline{t} = 2005$		GOM Deep, $\underline{t} = 2020$	
$\overline{x}$	4400	4400	700	700	700	700	700	700
ν	150	150	300	300	300	300	300	300
	ML-EST	FIT-URR	ML-EST	FIT-URR	ML-EST	FIT-URR	ML-EST	FIT-URR
$\hat{\beta}$	0.35	0.00	-0.30	-0.20	0.40	-0.30	0.50	-0.45
$\hat{\mu}$	2.72	3.95	6.61	6.08	2.42	6.09	2.18	5.61
	(0.21)	(0.26)	(1.54)	(1.30)	(0.21)	(1.14)	(0.16)	(0.79)
$\hat{\sigma}$	3.14	2.74	2.56	2.56	2.79	2.58	2.53	2.36
	(0.42)	(0.26)	(0.47)	(0.46)	(0.49)	(0.40)	(0.32)	(0.28)
Res[a]	1.10	12.75	47.99	42.83	8.01	40.93	3.31	38.34
	[0.39;2.01]	[6.51;19.80]	[34.73;58.63]	[31.20;52.42]	[5.61;10.57]	[31.33;49.53]	[2.28;4.45]	[29.86;46.13]
L[b]	-684.54	-692.38	-351.57	-351.65	-504.96	-505.93	-672.57	-680.53
LR[c]	15.68***		0.16		1.94		15.92***	

Note: Estimated standard errors are in parenthesis.

[a]Undiscovered resources in Gb. The upper row is the mean estimate, in brackets are the 10% and 90% quantiles.

[b]Log-likelihood evaluated at the parameter estimates. [c]Likelihood ratio statistic. *p<0.1. **p<0.05. ***p<0.01.

GOM Flat, which is consistent with the results from Tables 4.4-4.5. A positive value for β is generally associated with a mean parameter below and a standard deviation parameter above the respective figure from the observed sample. This can be seen, for example, by the estimates of the ML-EST specification for Norway, $\underline{t} = 1984$, which amount to $(\hat{\mu}, \hat{\sigma}) = (1.65, 3.29)$ in comparison to the values $(4.86, 2.01)$ from Table 4.3, while the conditional ML-estimates of the FIT-URR specification are in-between at $(\hat{\mu}, \hat{\sigma}) = (3.64, 2.90)$. Besides, we see that for GOM Flat the estimates for σ are mostly below the respective estimates for the other two regions. This is consistent with the empirical size distributions shown in Figure 4.6, where the kernel densities for GOM Flat do not reach out to very large field sizes, in contrast to what is the case for Norway. Comparing the giant fields in these two regions, the data show that there are only two oil fields with sizes above 500 Mb in GOM Flat,[66] while until 2022 there have been 17 giant discoveries in Norway.[67]

For GOM Flat, $\underline{t} = 1962$, the ML-EST specification implies a mean estimate for URR of 8.14 Gb, which is calculated by adding the mean estimate for undiscovered resources from Table 4.6 to the 4.62 Gb already discovered before 1962. The corresponding 10%- to 90%-quantile range is 7.29 - 9.06 Gb, which is below the known amount of 10.34 Gb discovered until 2020. For $\underline{t} = 1967$, the estimated URR even decreases somewhat compared to $\underline{t} = 1962$. Note that the likelihood ratio statistic for $\underline{t} = 1967$ becomes significant, so that this test would imply the rejection of the 10.34 Gb figure. Overall, we can conclude that the recoverable resources in GOM Flat are underestimated by the our estimation results.

Turning next to Norway, $\underline{t} = 1984$, Table 4.6 shows for the ML-EST specification a mean estimate of undiscovered resources of 8.11 Gb, and a 10%- to 90%-quantile range of 4.30 - 12.71 Gb. Adding this to the 22.90 Gb already discovered before 1984 yields a range of 27.20 - 35.61 Gb, which just contains 35.46 Gb, the total amount of oil discovered by 2022. Overall, the estimate can be judged as somewhat low but still reasonable if one assumes

[66]These are West Delta 030 and Bay Marchard 002 with 596 and 550 Mb of oil. If one includes the gas fields as well and measures size as the sum of oil and gas resources in barrels of oil equivalents, there are 7 giants with sizes in the range of 555-813 Mb of oil equivalents.

[67]The largest two are Stratfjord and Ekofisk with 3.67 Gb and 3.44 Gb of oil.

that no significant discoveries will be made in Norway from 2022 onwards. However, the NPD's estimates from 2021 imply that discoveries on the order of 15 Gb might still be made. Thus, in the FIT-URR specification for $\underline{t} = 1984$, the closest match to the 50 Gb figure for URR is achieved by setting $\hat{\beta} = 0.25$. Table 4.6 shows that this specification is also associated with wider interval estimates, implying for URR a 10%- to 90%-quantile range of 36.09 - 60.85 Gb.

For GOM Deep, the ML-EST specification for $\underline{t} = 2000$ yields a negative $\hat{\beta}$-coefficient of -0.3, implying mean undiscovered oil resources of 48 Gb, which upon adding to the 8 Gb discovered until 2000 gives an amount close to the external URR estimate of 53 Gb. However, for $\underline{t} = 2005$ and 2020, the undiscovered resource estimates are reduced considerably in the ML-EST specifications where now positive $\hat{\beta}$-coefficients of 0.4 and 0.5 and correspondingly much lower $\hat{\mu}$-coefficients are obtained. Adding to the quantiles for undiscovered resources the amounts already discovered by $\underline{t} = 2005$ or $\underline{t} = 2020$, which are 10.8 Gb or 12.75 Gb, respectively, we obtain for URR the 10%- to 90%-quantile ranges of 16.41 - 21.37 Gb or 15.03 - 17.2 Gb, respectively.

For both GOM Deep and Norway, we conclude that the ML-estimation of (β, μ, σ) mostly gives URR estimates that are substantially below what the respective agencies estimate. For both regions, the likelihood ratio statistics for the two earlier periods reported in Tables 4.6-4.7 are small, so that no statistically significant differences between the $\hat{\beta}$-coefficients from the FIT-URR and ML-EST specifications can be found. In contrast, for the most recent period for both regions the likelihood ratio statistics are clearly statistically significant, so that the FIT-URR specifications would be rejected based on this statistic, similar as for GOM Flat for $\underline{t} = 1967$. Note that for these most recent periods the ML-estimates of undiscovered resources become pretty small in comparison to what has already been discovered.

4.5.3 Simulation Results for the Overall Model

In this section, we examine how well the overall model from sect. 4.3 can predict the actual production and discovery rates. For this we conduct MC simulations as 10,000 runs of a simulation procedure based on Theorem 2. The simulations are based on the parameter values reported in the previous section for $(\overline{x}, \nu, \hat{\beta}, \hat{\mu}, \hat{\sigma})$, and the estimated covariance matrix for $(\hat{\mu}, \hat{\sigma})$. Additionally, the simulations require specification of all the other parameters introduced in sect. 4.3. Thus the next few paragraphs explain how these parameters are estimated / chosen. Afterwards, we present and discuss our main results from the simulations. For reasons of space we show in this section only the results from three specifications, namely a single ML-EST specification for each region, which is based on assumption 4b and the years $\underline{t} = 1962$, 1984, and 2000 for GOM Flat, Norway, and GOM Deep, respectively. The results are presented in Figures 4.8-4.10. In Appendix C, we also show the results from the corresponding FIT-URR specifications and the corresponding specifications based on assumption 4a.

For the ML-EST specifications, we want to stress that the only information which we use in the simulations and which is calculated from the full samples up to 2020 (or 2022 for Norway) are the durations between discovery and first oil (δ_0), and the parameter $\nu = 164$ for GOM Flat. The parameter δ_0 is always chosen as the sample median from the full samples, which amounts to approximately 4, 6, and 10 years for GOM Flat, GOM Deep, and Norway, respectively. Note that values of roughly similar magnitude could be

inferred by inspection of figure 1 shown by Mihalyi (2021). In all cases the parameters for the production profiles are based on IEA (2008) as explained in sect. 4.3.1. The upper row in Figure 4.1 shows the two profiles that we use for Norway. The panel "deepwater & non-giant" in Figure 4.1 shows what we use for fields in GOM Deep with sizes below 500 Mb. For giant fields in GOM Deep we use the corresponding profile where the decline rate is instead only 10.8%, as reported in Table 4.1. Although fields in GOM Flat are located offshore on the shelf, we find from comparing some actual production profiles with the ones generated from the parameters "shelf" vs. "onshore" that the more stretched-out profiles based on the onshore parameters fit much better, which are consequently used here.[68] Because there are only two giant oil fields in GOM Flat with sizes slightly above 500 Mb (596 and 550 Mb), we use the "onshore & non-giant" profile also for these two fields. Hence, the profile that we use for GOM Flat can be seen in the bottom-right panel of Figure 4.1.

For the FIT-URR specifications (shown in the Appendix C), we obviously use as "ex-post information" also the external estimates of the regional URR. To the extent that these numbers will turn out as correct, we thereby reduce the error about the area under the curves, while we can still evaluate how well the "shape" of the curves (i.e. the pattern how the discovery and production rates evolve over time) is predicted.

For the simulations based on assumption 4b, our preferred specifications, we also need to estimate the parameters a_0 and b_0. We plot in the right columns of Figures C5-C7 in Appendix C the historical trends for the number of exploration wells that were needed to discover a field, which is computed from the exploration well time series and the discovery times. Shown as the black line in the figures is a moving average of this trend that is computed using a time window of 10 years. By evaluating this function at the k discovery times within the sample period, we obtain the data $r_{\text{obs},i}^{-1}$, $i = 1, \ldots, k$, with which we estimate the parameters a_0 and b_0 from the regression (4.23). The resulting estimates $\hat{a}_0$ and $\hat{b}_0$ can be found in Table C1 in Appendix C.[69]

Each specification contains furthermore three scenarios for the trend functions $\Lambda(t)$, which we organize in the figures as three different columns. As explained in sect. 4.3.3, we estimate $\Lambda_{\text{b}}(t)$ by fitting a power function ct^d to the actual exploration well series. This provides only a single scenario for $\Lambda_{\text{b}}(t)$, which we always put under scenario 2 in the figures. To arrive at different $\Lambda_{\text{b}}(t)$-functions for scenarios 1 and 3, we simply multiply the estimated value of d by 2 or 1/2, respectively.[70]

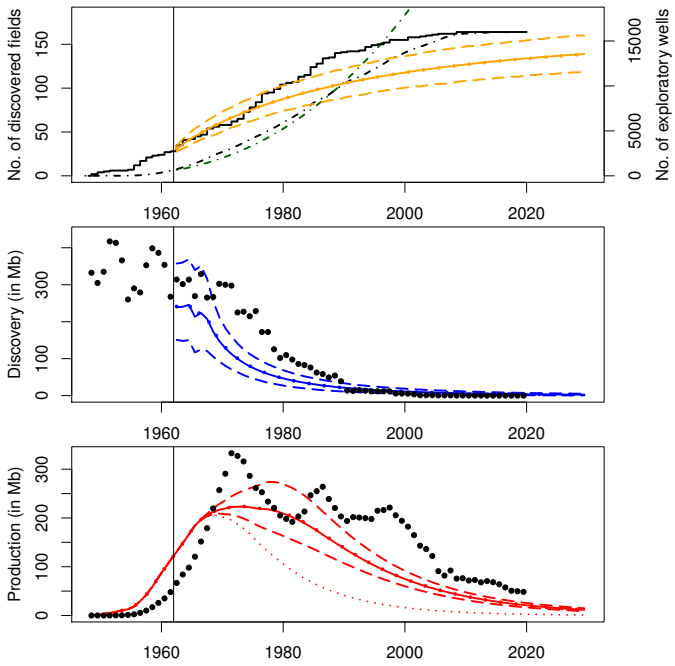
To explain how the figures for the simulation results are structured, we anticipate in Figure 4.7 the result of scenario 2 from Figure 4.8. First, observe that in each panel we

[68] This could be related to the fact that many of these fields are in very shallow water (the median water depth is 200ft), or to the fact that these fields were launched mostly in the 1950s-80s. The numbers published by IEA (2008) reveal that non-OPEC oil fields that were launched in the 1950s-80s had substantially lower decline rates than those that were launched in the 1990s-2000s (IEA (2008), p. 240).

[69] For the constant we have imposed the restriction $\hat{a}_0 \geq 1$ by setting $\hat{a}_0 = 1$ whenever $\hat{a}_0 < 1$ was obtained by OLS. We do this to make the denominator of (4.22) decline at least as fast as the "area" of the yet-to-be discovered fields. This parameter restriction is not necessary for assumption 4b as it is currently presented (see sect. 4.3.3), but it also should not affect the overall results too strongly.

[70] For the specifications based on assumption 4a, we employ three different parametric functions for $\Lambda_{\text{a}}(t)$: for scenario 1 we use an exponentially growing function bounded by 1, for scenario 2 we use a variant of the Gaussian cdf which is flexible to be asymmetric (see Brandt (2007), p. 3086), and for scenario 3 we use the cdf of the exponential distribution (i.e. a bounded growth curve). The parameters of these functions are estimated as illustrated in Figure 4.4.

Figure 4.7: Simulation results illustration (scenario 2 of Figure 4.8 for GOM Flat)



Note: The black vertical line marks the date $\underline{t}$. Colored curves are predictions while black refers to actuals. In the upper panel, the scale on the left pertains to the orange curves and the black step function, while the scale on the right pertains to the green and black dot-dashed lines. In the lower panel, the red dotted line shows the oil that would be produced if no new fields were discovered after time $\underline{t}$.

have marked the year $\underline{t}$ that separates the sample period from the projection period by drawing a black vertical line. The first row depicts as the green dot-dashed line the trend function $\Lambda_{\mathrm{b}}(t)$ for the cumulative number of exploration wells. The actual time series is depicted as the black dot-dashed line. Note that the scale for exploration wells is on the right-hand side of the panel. In contrast, the scale on the left-hand side of the panel pertains to the cumulative number of discovered fields, for which the actuals are shown by the black step function. The simulation results for this variable, which in mathematical terms is the stochastic process $(N_t)_{t\geq \underline{t}}$ as described in sect. 4.3.3, are shown in the form of its mean function (orange solid line) and its 5%- to 95%-quantile range (orange dashed lines).

The second row depicts in blue our simulation results for the smoothed discovery rate, which we compute using the function q_{disc} (eq. 4.4) with $\vartheta = 10$. Again the solid line is the mean function and the dashed lines show the 5%- to 95%-quantile range. The black dots show the actual smoothed discovery rates as calculated from the full available discovery history (until 2020 for GOM, until 2022 for Norway), and using q_{disc} with $\vartheta = 10$.

In the third row, we depict in red the production rate computed by using the production profiles q_{prod} (eq. 4.2) with parameters as explained above. Again the solid line, the dashed lines, and the black dots represent the mean function, the quantile range, and the actual yearly production time series, respectively. We have also added a red dotted line to show the amount of oil that would be produced if suddenly no new fields were discovered after time $\underline{t}$. Note that due to the assumed fixed durations between discovery and first oil production for each field, the red lines start to diverge only at $\underline{t} + \delta_0$ in each specification.

In each panel, we also show the asymptotic approximation to the mean function of the variable of interest. Be it the production rate, the discovery rate, or the cumulative number of discovered fields, this can be computed from (4.28-4.30) or related equations, plus either (4.26) or (4.32). The asymptotic approximations are indicated in the figures by colored dots that generally lie at or very close to the colored solid lines (the dots superimpose the solid lines in Figure 4.7). This confirms that the asymptotic approximations are nearly identical to the mean functions as computed from the MC simulations.

The graphical presentation of the results is complemented by Table 4.8 which reports for various specifications the goodness-of-fit of the mean discovery and production curves, and certain summary statistics for the mean production curves. The first two columns report the root mean squared prediction error (RMSPE), after division by the mean of the respective actuals over the prediction years to facilitate comparison among the regions and specifications. The summary statistics in the other columns pertain to the "shape" of the mean production curves, similar as the summary statistics in Table 4.1 pertain to the shape of the field-level production profiles. Thus, Table 4.8 allows to compare the shape of the production curves resulting from assumptions 4b vs. 4a, ML-EST vs. FIT-URR, and from two different scenarios for $\Lambda(t)$ in each case.

Turning now to the results, we start with Figure 4.8 for GOM Flat. The figure shows that the quantile ranges for both discovery and production are too low compared to the actuals in most years. This is due to the fact that the ML-EST specification implies for the integral under the discovery and production curves a range of 7.29 - 9.06 Gb, while the observed total discoveries until 2020 are 10.34 Gb. Nonetheless, except for this scale difference the pattern how the discovery rate declines over time is matched to a reasonable degree by our forecast. To eliminate the effect from having underpredicted the total resources, we consider for comparison the corresponding FIT-URR specification, which we show in Figure C12 in Appendix C. Note that here scenarios 2-3 show a very good correspondence of the projected discovery rate and the actuals, and that the actual production lies within the quantiles for many years. Besides, when using assumption 4a with a Gaussian trend function as shown in scenario 2 of Figure C13 in Appendix C, we have by chance obtained an almost perfect match to the number of discovered fields. Here the resulting discovery rate is also matched almost perfectly by the forecast, while for the production rate there remains some discrepancy.

The strongest mismatch of our forecasts for GOM Flat is the unanticipated stagnation and rise in production during the 1990s. This is underpredicted in Figure 4.8 and Figures C12-C13, where a declining trend is predicted instead. In contrast to our production forecasts (the red solid lines) which always show a single local maximum, the actual production curve oscillates and has three local maxima during 1962-2000. This indicates that several other factors (e.g. economic, regulatory, environmental or technology-related factors) that are omitted in our model play a role for the discovery and production dynamics here. For example, the duration between discovery and first oil production may not be constant as assumed, but instead can systematically change over time as a function of oil prices, regulatory factors, or due to experience gained from past development projects. Similar factors may dynamically affect the initial production build-up phase for many fields. Alternatively, the observed rise in production during the 1990s could also come from secondary and tertiary enhanced oil recovery projects applied to several of the older fields that at this time were approaching the end of their production years. All of this implies that the mismatch arises because field-level production is to some extent

Table 4.8: Summary statistics for the simulation results

	Specification	Scenario	RMSPE disc.	RMSPE prod.	% of URR prod. at peak	% of URR prod. until peak	Number of years at plateau[a]	Rate of increase pre-peak (%)	Rate of decline phase 1[b] (%)	Rate of decline phase 2[c] (%)	Rate of decline phase 3[d] (%)
GOM Flat, $\underline{t} = 1962$	4b,ML-EST	1	0.71	0.39	3.0	46	17	21.7	2.3	6.6	9.2
	4b,ML-EST	2	0.73	0.39	2.7	33	17	25.3	1.5	5.2	6.5
	4b,FIT-URR	1	0.89	0.38	3.3	47	12	21.4	2.6	7.3	9.7
	4b,FIT-URR	2	0.47	0.29	2.8	43	15	20.8	2.1	5.7	7.5
	4a,ML-EST	1	0.67	0.36	2.6	22	17	30.7	1.3	3.1	8.7
	4a,ML-EST	2	0.58	0.36	2.6	28	21	27.7	1.1	4.7	8.8
	4a,FIT-URR	1	0.42	0.23	2.2	47	29	17.2	1.4	4.4	9.3
	4a,FIT-URR	2	0.18	0.24	2.5	50	22	17.7	1.9	5.2	8.8
Norway, $\underline{t} = 1984$	4b,ML-EST	2	0.61	0.16	3.4	42	12	16.0	2.2	6.7	10.6
	4b,ML-EST	3	0.57	0.14	3.2	42	12	15.8	2.8	5.5	8.2
	4b,FIT-URR	2	1.76	0.52	2.7	50	19	12.6	1.9	6.4	9.7
	4b,FIT-URR	3	1.04	0.27	2.2	27	19	16.0	1.2	3.1	5.3
	4a,ML-EST	2	0.63	0.17	3.5	42	11	16.2	1.8	7.4	10.5
	4a,ML-EST	3	0.56	0.14	3.2	42	12	15.8	2.7	6.0	8.2
	4a,FIT-URR	2	1.82	0.56	2.9	44	15	14.0	1.6	6.3	9.6
	4a,FIT-URR	3	1.06	0.29	2.2	27	18	16.0	1.1	3.0	5.1
GOM Deep, $\underline{t} = 2000$	4b,ML-EST	1	7.18	0.79	3.5	52	13	15.3	2.7	8.0	38.9
	4b,ML-EST	3	2.11	0.30	1.3	24	26	14.5	1.0	1.7	2.0
GOM Deep, $\underline{t} = 2005$	4b,ML-EST	1	2.32	0.26	3.4	56	15	15.3	2.5	8.5	38.5
	4b,ML-EST	3	1.42	0.25	2.9	34	15	18.3	1.7	5.4	6.8
	4b,FIT-URR	1	10.50	0.35	3.2	53	13	14.0	2.1	7.6	33.4
	4b,FIT-URR	3	3.90	0.26	1.4	29	25	13.5	1.1	2.0	2.4
	4a,ML-EST	1	2.42	0.25	3.2	62	18	14.4	2.8	9.7	59.7
	4a,ML-EST	3	1.77	0.26	3.0	46	16	16.3	1.9	6.5	13.4
	4a,FIT-URR	1	13.83	0.30	4.4	55	7	14.7	3.6	10.9	25.0
	4a,FIT-URR	3	5.56	0.29	1.8	50	38	11.3	0.6	8.7	14.2

Note: The division of the decline phase into three separate phases is based on IEA (2008, p. 235).
[a]Defined as the period during which production is more than 85% of peak level. [b]Defined as peak year until end of plateau.
[c]Defined as end of phase 1 until production reaches 50% of peak level. [d]Defined as end of phase 2 until the last year from our calculations.

Figure 4.8: Simulation results for GOM Flat ($\underline{t} = 1962$, ML-EST, assumption 4b)

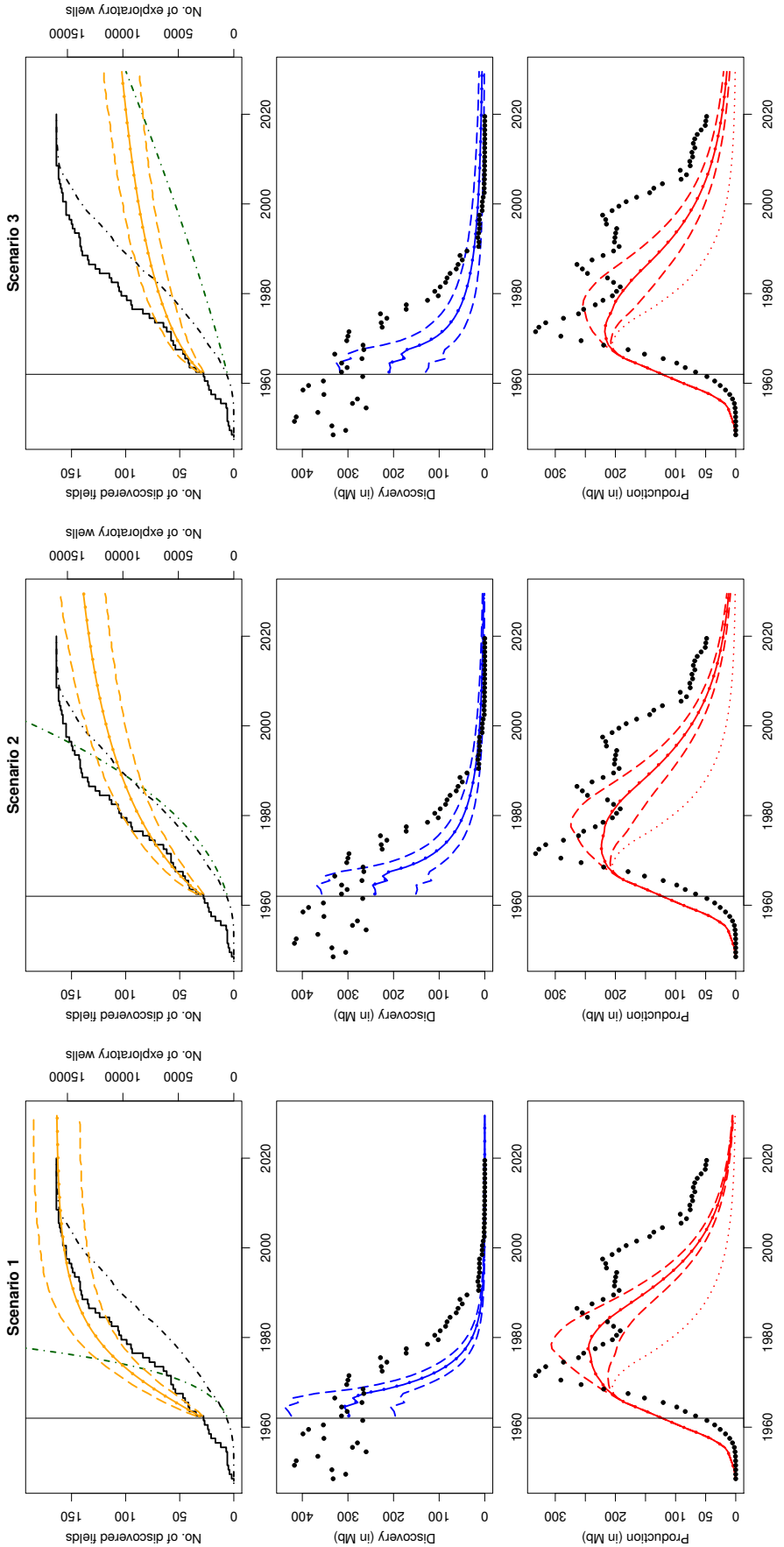


Figure 4.9: Simulation results for Norway ($\underline{t} = 1984$, ML-EST, assumption 4b)

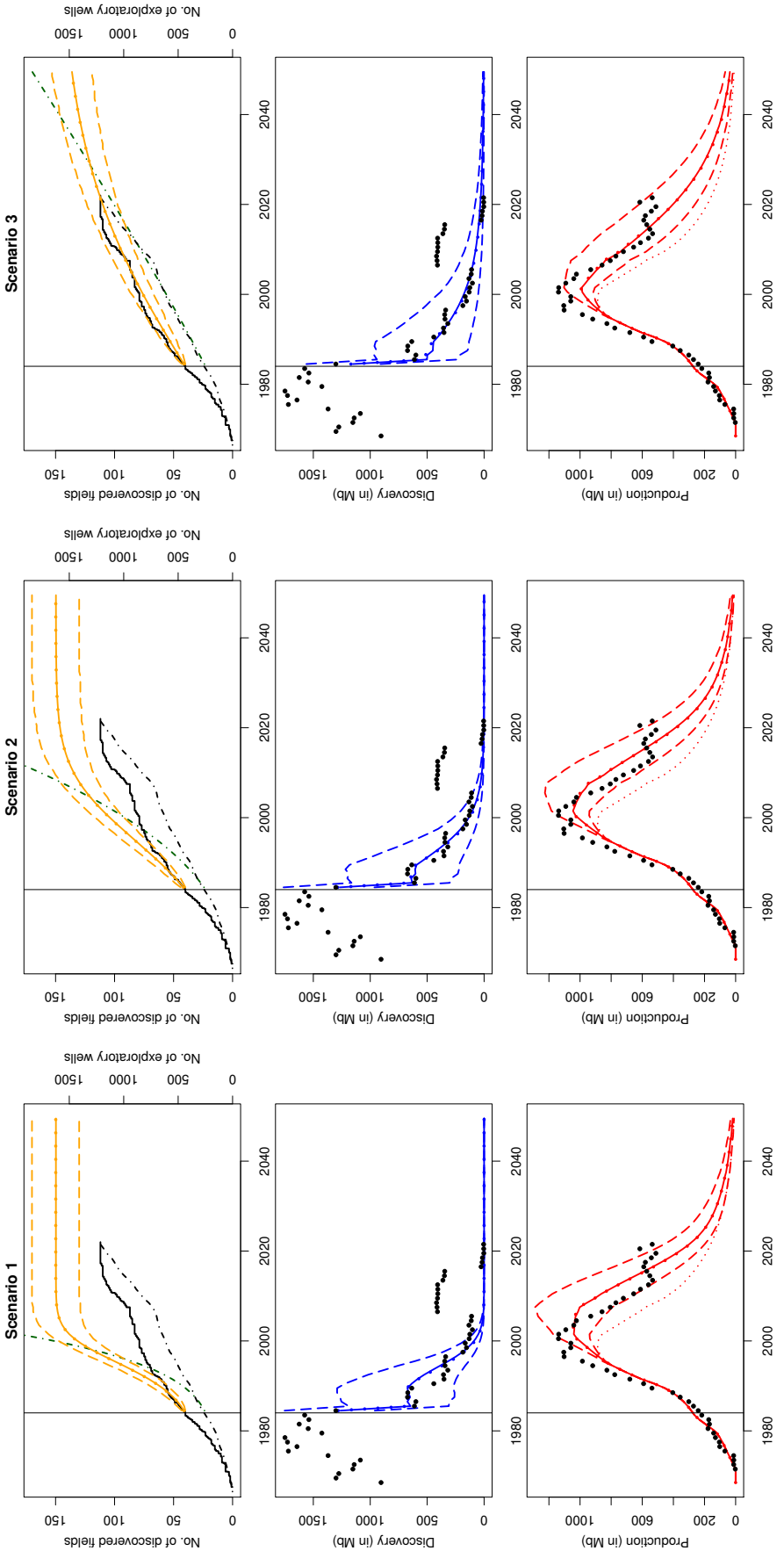
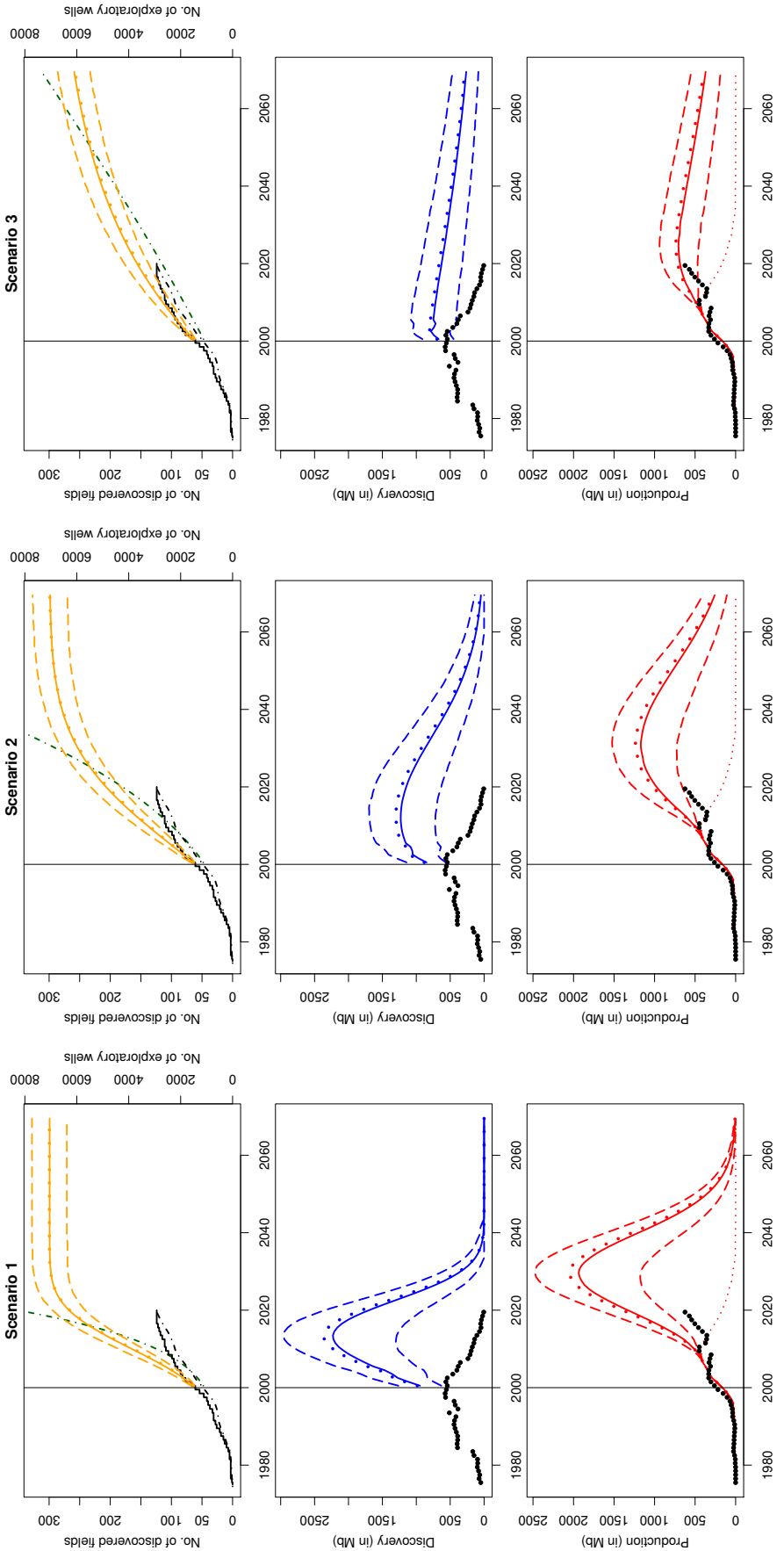


Figure 4.10: Simulation results for GOM Deep ($\underline{t} = 2000$, ML-EST, assumption 4b)



dynamically influenced by several factors instead of only being a function of time since discovery.

For Norway, Figure 4.9 shows a fit of the forecasts to the actuals that is overall good. Until 2005, the mean forecast of the smoothed discovery rate matches well to the actuals. The increased discovery rate during 2005-15 is largely driven by the 2010-11 discoveries of the giant fields Johan Sverdrup and Johan Castberg, which contain about 2.6 Gb and 560 Mb of oil, respectively. That such large discoveries occur after the discovery rate has declined to rather low levels is not foreseen by the forecasts. On the other hand, the increase and decline of production around the peak year 2000 is well anticipated by the forecasts. Also the suspension of the declining trend for production in the 2010s is well within the quantiles. The actuals exit the quantile range only since 2020, after the Johan Sverdrup field has started to produce in late 2019. Let us compare Figure 4.9 to the corresponding FIT-URR specification (with a mean of additional 17 Gb of oil) shown in Figure C14 in Appendix C. In scenario 3, the scenario with the best fit regarding exploration wells, a large discovery around 2010 just like Johan Sverdrup and Johan Castberg is predicted. Turning to the production forecast, the figure shows actuals that lie between the predicted lower quantile and mean even during the 2010s. Comparing the two figures for Norway shows that the additional 17 Gb of oil do not substantially delay the predicted year of peak production, but considerably mitigate the post-peak decline. This is confirmed in Table 4.8 numerically: while peak production occurs in all specifications before or when 50% of the URR is extracted, the plateau lengths are increased and the decline rates are reduced when comparing the FIT-URR to the ML-EST specifications.

Let us provide possible explanations for why the latest large discoveries around 2010 are underpredicted in Figure 4.9, which will soon after 2021 also cause the production forecast to underpredict the actual production. One important issue here pertains to the regulatory access to the different parts of the Norwegian continental shelf which needs to granted by the regulator (i.e. the NPD) before oil exploration can take place. This happened at different times for the different parts of the region: while exploring for petroleum in the North Sea was permitted already since 1965, exploring in the Norwegian and Barents Sea was permitted only since 1980, and even nowadays access to vast areas of the Barents Sea is not yet permitted (NPD (2021), p. 3). A more "natural" discovery history as consistent with the axioms of the discovery process model has thus been prevented. In particular, these axioms require that all fields are in principle available for discovery at each time while only size governs the probabilities to observe a particular discovery order (see assumption 2 in sect. 4.3.2). Clearly, this assumption is violated by the differential access to the different parts of the Norwegian continental shelf, which has introduced additional bias into our parameter estimates and thus into the discovery and production forecasts.

Since we have not treated the discovery sequence of the Barents and the Norwegian Sea as separate samples but instead pooled the fields together with the North Sea fields, their resource potential is likely underestimated in the results shown in Tables 4.6-4.7 and Figure 4.9. For example, the giant field Johan Castberg that was discovered in 2010 is located in the Barents Sea. In contrast, the huge Johan Sverdrup field is located in the North Sea where access was possible since 1965, so that the explanation regarding the regulatory access does not apply to Johan Sverdrup. Indeed, the field was missed by 200 meters by an exploration well in 1967 (Lillestøl and Sinding-Larsen (2017), p. 69). Thus we need to find other reasons for why our model underpredicts a large discovery like Johan

Sverdrup. Note for this that the late discovery of Johan Sverdrup is fully consistent with the probabilistic nature of the discovery process model which permits that huge discoveries can occur late in the discovery process, albeit with low probabilities. A possible reason for why our predictions imply that the probability for this event is too low could be because the total number of fields has been chosen too low with $\nu = 150$. This is in line with results by Lillestøl and Sinding-Larsen (2017; 2018) who estimate a creaming discovery process model based on Meisner and Demirmen (1981) for Norwegian oil and gas fields. Lillestøl and Sinding-Larsen (2017, p. 79) find that allowing for a sufficiently large number of new fields is necessary so that their resource estimate matches the one from the NPD, and in this case the authors calculate that the chance to find another field like Johan Sverdrup is 12%.

Turning now to GOM Deep, Figure 4.10 shows the results of the ML-EST specification based on $\underline{t} = 2000$ and assumption 4b. In each scenario the huge additional resources (of 48 Gb) are discovered at rates drastically above the actual rate that instead declines during 2005-20, showing that such a large resource potential is at odds with the actual discovery record until 2020. As the resource potential of the FIT-URR specifications is of similar magnitude (see Table 4.7), this casts some doubt on the large undiscovered resource estimates as reported in BOEM (2017). Thus, the deviation of our forecasts in Figure 4.10 from the actuals is due to greatly overestimated resources. It remains the question why the ML-estimation procedure for $\underline{t} = 2000$ has given such an overestimated resource potential. From Figure C10 in Appendix C one can see that the likelihood profile from which $\hat{\beta}$ was determined is very flat. In fact, it is not possible to reject with a likelihood ratio test at the 10% significance level any β-values in the range of $[-0.5, 0.65]$. The higher values in that range would imply a much lower resource potential.[71] The reason for the flat likelihood profile is probably the lack of a clear declining trend in the discovery size per discovered field before the year 2000 (see Table 4.4, or the middle panel of Figure C7 in Appendix C). As discussed at the end of section 4.5.1, this is probably because there is instead a clear depth bias in the discovery sequence which overshadows the size bias.

Among all the specifications we estimated for GOM Deep, the ML-EST specification for $\underline{t} = 2005$ with an implicit total resource estimate of 19 Gb fits best to the actuals. This is shown in Figure C16 in Appendix C, where scenario 3 provides a match with the actuals that is not too bad. Even here the predictions for the discovery rate are persistently above the actuals. Potentially, this mismatch could also arise if the published reserve estimates pertaining to the more recent discoveries are underestimations of the "true" remaining reserves. The could be either since several pools discovered before 2020 are missing in the reserve reports, e.g. because they are not yet assessed in terms of the profitability of recovery, or due to future reserve growth. If substantially increased field sizes are reported in future reserve publications, especially for the more recent discoveries since reserve growth more strongly affects fields in the initial years after discovery (see, e.g., figure 1 in Schmoker and Klett (2000)), the decline of the discovery rate during 2005-20 might not be as pronounced as shown in the figures.

Despite the general mismatch of the forecasts in Figure 4.10 when compared to the actuals, scenario 3 of the figure provides a nice example that illustrate some conditions under which the model produces curves without evident peaking, i.e. where the production

[71] Also the rather high value $\nu = 300$ (or 250) was chosen for the total number of fields, which has a large impact on the resource estimate for $\hat{\beta} = -0.3$ (see the right panel of Figure C10).

and discovery rates are rather flat. For scenario 3 of Figure 4.10 we have specified a nearly linear trend function for exploration wells, with the result that the pace at which new fields are discovered (as shown by the orange lines) gradually slows down over time. Combined with the negative value $\hat{\beta} = -0.3$, which implies that successive fields become larger on average, this yields extended plateaus for both the discovery and the production rate. Table 4.8 shows that in scenario 3 of Figure 4.10 the production is at a plateau for more than 25 years and that the decline rates are very low. Thus, we have obtained here a constellation of parameters where discovery and production clearly deviates from a bell-shaped or hump-shaped curve. In contrast, scenario 1 shows a strongly increasing rate of exploration effort (that is probably highly unrealistic logistically or in terms of economic costs), resulting in a fast discovery of the remaining fields, and in discovery and production curves with a pronounced peak and a sharp decline side (see Table 4.8).[72]

Overall, the logic of the model and our empirical results suggest that flat profiles emerge only for the case of non-positive or "very low" values of $\hat{\beta}$. To illustrate this, consider the FIT-URR specifications for Norway as an example, shown in Figures C14-C15 in Appendix C. Here the amount of undiscovered resources is also quite large but $\hat{\beta} = 0.25$. The figures show that the discovery and production curves are again rather hump- or bell-shaped than flat (though this conceptual distinction is, of course, a gradual one), and that the additional resources do not prevent production from declining but only mitigate the rate of decline.

As a side note to the results for GOM Deep, note that average water depth shows an increasing trend with discovery order (see Tables 4.4-4.5), and that water depth can range up to 9000 ft for these deepwater fields. Hence, even if the large undiscovered resources as estimated in BOEM (2017) are to be discovered in the form of many large but at the same time deeper fields, the energy required to extract oil from these fields will be substantially higher. As the quantity that is relevant to the economy is the net energy that remains after deducting the energy required for extraction, a hypothetical curve for net energy for this case could be substantially below the oil production curves shown in Figure 4.10.

4.6 Conclusion

Let us conclude by summarizing and commenting on the major takeaways from our empirical analysis. At first, the descriptive analysis in sect. 4.5.1 has established with simple methods that there is a size bias for oil field discoveries in GOM Flat and Norway. For GOM Deep the size bias is less evident and is only present when analyzing all discoveries until 2020. The finding that larger oil fields tend to be found first is not only in alignment with "industry folklore" but also with similar findings in the literature for the North Sea (e.g., Smith and Ward (1981); Michel (2011)), and also for other oil plays (e.g., Lee (2008); Chen and Osadetz (2009)). Since the size bias is the key axiom of the size-biased sampling discovery process model, the model seems in principle suitable based on the descriptive analysis. Because the axioms as stated in sect. 4.3.2 also imply that size is

[72] To bring the point home we show in Figure C17 in Appendix C also the result from a specification based on assumption 4a (for $\underline{t} = 2005$ and with the parameters from the FIT-URR specification in Table 4.7). For scenario 3 we have extrapolated a bounded growth curve for the cumulative number of discovered fields, with the result that the production and discovery rates have flat profiles until all remaining fields are found. (To fit the bounded growth curve, however, we had to add the arbitrary assumption that all fields are found by 2050.)

the only variable that influences the discovery order (and that the residual influence is "chance"), a violation in the form of other factors with strong influence on the discovery order can, of course, introduce bias into the estimates derived from the model. In sect. 4.5.1 we have found that water depth is another variable that influences the discovery order. In particular, we have found for each region that there is a clear tendency of increasing water depth, i.e. a depth bias, which is in line with what is known for the deep water plays in the GOM (Burgess et al. (2021), p. 13; Managi et al. (2005)). This finding is not fully surprising since deep water technologies such as tension leg platforms or subsea completions have become commercial only since the 1980s (Managi et al. (2005), pp. 622f.). Lastly, sect. 4.5.1 has found for all three regions that the lognormal distribution provides a fairly good approximation to the empirical field-size distribution and seems better suited than the Pareto distribution.

Secondly, we have reported in sect. 4.5.2 the undiscovered resource and URR estimates as estimated via maximum likelihood from the discovery process model. For most of the selected years for GOM Flat and Norway the ML-estimation has returned point estimates for URR which are below the resources known with hindsight as of 2022. When all data until the most recent year are used, the remaining resource estimates are often fairly small compared to the resources that are already discovered. Besides, for Norway and for GOM Deep (for $\underline{t} = 2005, 2020$), the resulting URR estimates are substantially below the point estimates published by the official agencies (i.e. the NPD and the BOEM). Of course, the "true" URRs in the regions GOM Deep and Norway which are not yet fully explored are unknown quantities associated with considerable uncertainty. Our estimates for GOM Flat provide an example where the ML-estimation has to some degree underestimated the "true" URR (based on total field sizes discovered by 2020).

Thirdly, as presented in sect. 4.5.3, the main purpose of this paper was to explore which curves emerge for the discovery and production rates of the overall bottom-up model, and to analyze how well these curves can predict the actuals. For GOM Flat and Norway, using discovery sequences until the years 1962 and 1984, respectively, has yielded discovery and production curves which to a reasonable degree match to the pattern of the actuals, even quite far out-of-sample. Possible reasons for the remaining mismatch between the discovery and production curves and the actuals for GOM Flat and Norway have been discussed in sect. 4.5.3. For GOM Deep our predictions are overall less conclusive, driven by large discrepancies between the estimated amounts of undiscovered resources when using discovery sequences of different lengths and in comparison to the estimates from BOEM (2017). Here the estimation based on the discovery sequence until 2000 has returned a discoverability parameter β that is negative with a consequentially large undiscovered resource estimate. This has happened probably because of the lack of a clear pattern of size bias within the discovery sequence before 2000, where instead there is a clear bias due to water depth. In the corresponding predictions, the simulated discovery rates drastically overpredicts the actual discovery rate.

Despite the mismatch of these specifications for GOM Deep when compared to the actual discovery rate, the results have given interesting insights into the conditions under which curves that are "rather flat" can emerge. When the negative β parameter is combined with a scenario where there is a gradual decline in the pace at which new fields are discovered, we obtain curves where the production rate is rather flat instead of following a hump- or bell-shaped curve. Overall, though, when the discoverability parameter β is "far enough" above zero, our results indicate that roughly hump- or bell-shaped curves emerge and

that oil production peaks when - or mostly before - half of the oil is extracted. Campbell and Laherrère (1998, p. 80) have claimed in support of the Hubbert model that "adding the output of fields of various sizes and ages [...] usually yields a bell-shaped production curve for the region as a whole". The bottom-up model presented in this paper provides a mathematical framework to examine this phenomenon. Of course, several similar and differing models or empirical calculations have been brought forward as an explanation for this phenomenon (aside from the references mentioned in sections 4.1-4.2, see for example Miller and Sorrell (2014), p. 14; Stark (2008); Bentley (2002); Michaelides (2017); Mohr and Evans (2008); Bardi (2005); Kaufmann and Cleveland (2001)). Thus, the framework described in this paper should be regarded as complementary.

The nature of the bottom-up approach is such that it is allows for various extensions (Jakobsson (2012), p. 71). It can also serve as a theoretical guidepost to discover new research questions (Jakobsson et al. (2014), p. 121). For example, we have specified for the oil production profiles that the duration between discovery and first oil production is constant. Additional research could further investigate this issue, as in Mihalyi (2021), and introduce time dependent or field-size dependent durations into the model (see also the discussion at the end of sect. 4.3.4). Moreover, with the constant parameter for the "empty area" (b_0) we have assumed that the success rate of exploration wells diminishes after each discovery. In this regard, future research could examine whether and to which extent the information gains that accumulate with exploration activity narrow down the area that is searched for new oil fields. Based on this the assumption of a constant b_0 and thus of monotonously declining success rates could be relaxed. Also, besides the number of exploration wells there are other proxies for exploration effort, such as the cumulative meters drilled by exploration wells. This measure can account for variation in drilling depth and also for horizontal drilling that has become widespread due to technical advances.

Furthermore, since our estimation procedure has shown the tendency to underestimate the regional URR, future research could instead apply the multivariate discovery process model as in Nair and Wang (1989), Lee (2008, pp. 40ff.), or Chen and Osadetz (2009). Here the successive sampling probabilities from assumption 2 are not only a function of size (recoverable oil), but instead of a multiplicative index of several jointly distributed lognormal variables. This would allow not only to include size into the estimation procedure but also water depth, or to break down the size variable according to the reservoir equation into the variables areal extent, net pay, et cetera (see Chen and Osadetz (2009), p. 107).

Finally, another avenue would be to use a different discovery process model for the estimation of the undiscovered resources and the prediction of the undiscovered field sizes. In particular, one could use the "creaming" discovery process model introduced by Meisner and Demirmen (1981) and recently used by Lillestøl and Sinding-Larsen (2017). Here the size bias phenomenon is introduced by a mean parameter (the "creaming factor") that diminishes with the successive discoveries where the field sizes are distributed as independent lognormal random variables, which makes the whole estimation procedure computationally much simpler. Lillestøl and Sinding-Larsen (2017; 2018) use an exponential functional form for the creaming factor in their empirical application to Norwegian oil and gas fields, and obtain a good fit to their dataset and to the resource estimates published by the NPD. A particular nice feature of the creaming model is that there is no need to specify the parameter for the total number of fields (n or ν) prior to the estimation

of the other parameters, so that misspecification of this parameter cannot interfere with the parameter estimation. Besides, the creaming model is amenable to a full Bayesian estimation, see Lillestøl and Sinding-Larsen (2018). Although the model based on successive size-biased sampling from a finite population is conceptually more satisfying, our results suggest that it can have poor predictive power for certain discovery sequences, and it certainly involves the problem of specifying the total number of fields because the estimation thereof is fraught with problems. In other words, although the simpler creaming model may not be "the best possible explanation of the exploratory behavior itself" (Lillestøl and Sinding-Larsen (2017), p. 82) it may give reasonable predictions and is computationally simpler. Therefore, future research could work on combining the creaming model with the other parts of the overall model (field-level production, exploration success and discovery times) presented in this paper.

5 A Wavelet Analysis of the German Wholesale Electricity Market Using Daily Data

5.1 Introduction

Understanding electricity market dynamics and their evolution over time is important for market participants such as power generators, grid operators, regulators, and industrial electricity consumers. Electricity market dynamics can be subject to further changes in the future if sector coupling raises the share of electricity in total energy consumption. Sector coupling or electrification refers to the replacement of non-electric energy (mostly fossil) used in sectors such as manufacturing, transport, and heating by electricity, preferably generated from renewable energy sources (i.e., wind, solar, biomass, geothermal). This is regarded as a cornerstone of current plans (e.g by the EU) to decarbonize the energy system (Van Nuffel et al. (2018); EU (2018); EU (2020)).

This paper examines the determinants of wholesale electricity prices with methods that can account for their cyclical as well as possibly nonstationary behavior. While for households and most smaller industrial customers retail prices as charged by the electricity providers are of primary interest, retail prices are dynamic only over longer cycles due to contractual fixation and are also subject to data availability issues. Therefore, this paper examines only wholesale electricity prices. In light of the planned rapid expansion of variable renewable electricity generation capacity (VRE, including wind and solar) and the plans for sector coupling, important insights can arise from investigating how an increasing VRE share interacts with the supply from conventional power plants, whose cost are in large part a function of fuel input prices (e.g., coal, gas, uranium), to determine wholesale electricity prices. Since each of these time series (VRE generation, fuel prices, electricity prices) moves in characteristic periodic cycles that are composed of multiple frequencies, a fundamental issue here is that the interactions among the different variables can differ substantially by frequency. As a further complication the periodic cycles and their interactions can also change over time. Wavelet methods are best adapted to disentangle these interactions by frequency and trace their evolution over time.

In this paper, I conduct a wavelet and cross-wavelet analysis using daily German electricity market data and daily prices of coal and gas for the period 2015-2023. The German electricity market data include time series on the wholesale electricity price, total electricity demand, and electricity generation by generation technology from which I obtain a time series on VRE generation. Applying bi- and multivariate wavelet methods allows me to examine the co-movement of electricity demand, VRE generation, and fossil input prices with electricity prices as a function of frequency and time. This is an advantage to spectral analysis methods (which operate only in the frequency domain) and to more traditional time domain methods (e.g., vector autoregression, cointegration).

Wavelet analysis is a commonly used method in economics already for some while. Some recent applications are, among others, Rua and Nunes (2009), Alvarez-Ramirez et al. (2012), Aguiar-Conraria et al. (2012), Verona (2016), Flor and Klarl (2017), Aguiar-Conraria et al. (2020), Krüger and Neugart (2020), and Krüger (2021). There is also a growing literature where wavelets are used to study topics in energy economics, including for example Sousa et al. (2014), Papaioannou et al. (2015), Aguiar-Conraria et al. (2018), Aguiar-Conraria et al. (2021).

This paper proceeds as follows. Sect. 5.2 reviews the related literature on electricity price determinants, including both the related literature where more traditional statistical methods are used and where wavelets methods are used. In sect. 5.3, I explain the methodology of wavelet and cross-wavelet analysis. Sect. 5.4 presents the time series data, and sect. 5.5 present the results of the wavelet analysis. The results section is structured into three subsections for the results from univariate, bivariate, and multivariate wavelet analysis, respectively. Finally, sect. 5.6 concludes.

5.2 Literature Review

There is a large literature studying the relations between renewable electricity generation, total electricity demand, and wholesale electricity prices. In this literature, many papers find that marginal increases of renewable electricity generation tend to reduce wholesale electricity prices (e.g., Sensfuß et al. (2008); Nicolosi and Fürsch (2009); Würzburg et al. (2013); Ketterer (2014); Cludius et al. (2014); Benhmad and Percebois (2018); de Lagarde and Lantz (2018); Sapio (2019); Maciejowska (2020); Cevik and Ninomiya (2022)). Since Sensfuß et al. (2008) this effect has been referred to as the merit order effect. The merit order refers to an ordering of the different power producers in terms of increasing marginal cost, yielding a short-term marginal cost curve. Intersecting the curve with an electricity demand curve that is inelastic during a short time period (e.g. an hour) determines the marginal power producer and the market price for this time period. Since renewable energy providers have negligible marginal cost (but instead fixed cost), when renewable energy production is high (or demand is low) only the conventional power plants with comparatively lower marginal cost are producing, and thus prices are reduced.

In this way, several papers identify the residual demand as a key determinant of electricity prices (Nicolosi and Fürsch (2009); Cludius et al. (2014); Wozabal et al. (2016); Kyritsis et al. (2017)). This quantity is defined as total electricity demand minus renewable generation, which therefore is the quantity that needs to be supplied by conventional power plants. For the German electricity market, one can observe a clear-cut positive correlation between the residual demand and the wholesale electricity price (Nicolosi and Fürsch (2009); Cludius et al. (2014)).

Moreover, several papers examine how renewable electricity generation impacts the volatility of wholesale electricity prices, overall finding that wind power increases price volatility while solar power tends to reduce it (Ketterer (2014); Kyritsis et al. (2017); Rintamäki et al. (2017); Maniatis and Milonas (2022)). Other papers use quantile regression to analyze which factors impact the distribution quantiles of wholesale electricity prices (Do et al. (2019); Sapio (2019); Maciejowska (2020)).

For empirical examinations of electricity prices as a function of renewable generation share, instead of analyzing marginal effects on wholesale electricity prices based on the merit order, more relevant for households and many industrial customers might be long-run or cross-country analyses of retail (i.e. end-user) electricity price levels. In Germany, retail prices include, in addition to passing on wholesale prices, distribution cost and profit margins for electric utilities, grid fees as compensation for transmission grid operators, electricity taxes and concessions, as well as certain levies. The biggest of these levies in Germany was the so-called "EEG-Umlage" introduced in the year 2000, which was levied to compensate the grid operators, who are obliged to remunerate any renewable electricity generation with fixed feed-in tariffs or a market premium to the wholesale price (§19 of

EEG, German Renewable Energy Law). Since the 1st of July 2022, however, the levy is no longer payable by end-users but instead paid for by the German government.[73]

There are few studies examining retail electricity prices, probably due to data availability issues. Two recent studies are Klopcic et al. (2022) and Hannesson (2024). For example, Klopcic et al. (2022, p. 5) calculate that German retail electricity prices for households increased from 23.8 ct/kWh in 2010 to 27.4 ct/kWh in 2013, but then remained roughly constant until 2019. Hannesson (2024) examines correlations between VRE share and retail electricity price levels for households and industry both via cross-sectional and time series regressions. The cross-sectional analysis reveals that there is a positive relation between the two variables across countries, and via individual time series regressions for 28 countries Hannesson (2024) finds that approximately half of the countries have positive and statistically significant regression coefficients. However, Hannesson (2024) also finds that retail electricity prices have declined in Denmark after reaching a share of above 40% for VRE generation, and that in Germany and Australia retail prices have stopped to increase since about 2013 while the VRE generation share has continued to rise.

Turning now to the literature that examines the impact of fuel cost of conventional power plants on wholesale electricity prices, this literature is embedded in the literature on the long-run co-movements among the prices of different energy commodities. The methods commonly employed here are cointegration testing and vector error correction modeling (VECM). Earlier studies that analyze the co-movement of energy and electricity prices include, among others, Emery and Liu (2002), Asche et al. (2006), Bunn and Fezzi (2008), and De Jong and Schneider (2009). For instance, Bunn and Fezzi (2008) estimate a VECM with daily spot prices of electricity, gas and carbon allowances for Germany and the UK for the period of 2005 until mid 2006. Key results by Bunn and Fezzi (2008), which apply to both countries, are the existence of a single cointegration relation between the three price series, and the result that the error correction occurs almost entirely by the electricity price, i.e. that the electricity price follows gas and carbon prices and not vice versa, as is consistent with role of these prices as cost fundamentals.

Frydenberg et al. (2014) examine the existence of cointegration between front-month future prices of coal, oil, gas, and electricity for Germany, the UK, and the Nordic countries, using daily data with coverage from 2006-2012. They find that electricity prices in all three country groups are cointegrated with coal prices, and that UK electricity prices are also cointegrated with gas prices. Moreover, Papaioannou et al. (2018) examine the Greek electricity market and find similar results as Bunn and Fezzi (2008). Papaioannou et al. (2018) use daily data from 2007-2014 on the Greek electricity spot price, the prices of Brent oil, gas futures, and carbon allowances, as well as Greek lignite prices. Testing for cointegration and estimating a VECM, they find a single cointegration relation between all prices, where also here solely the adjustment of the electricity price towards equilibrium is statistically significant and quantitatively relevant. Computing also a variance decomposition and impulse response functions, Papaioannou et al. (2018) identify gas futures prices and carbon prices as the two main drivers of the Greek electricity price, while the contributions of Brent oil and lignite prices are found to be negligible.

Cointegration analysis and vector autoregressions are time domain techniques. In contrasts, wavelet methods allow analysis in both time and frequency domain. Wavelet analysis is a commonly used method in economics already for some while. Some early

[73] See https://www.bundesregierung.de/breg-de/themen/tipps-fuer-verbraucher/eeg-umlage-faellt-weg-2011728

applications in economics and finance are Ramsey and Lampart (1998b; 1998a), Gençay et al. (2005), Gallegati et al. (2011), which all rely on the discrete wavelet transform. Later papers started using the continuous wavelet transform (e.g., Rua and Nunes (2009); Alvarez-Ramirez et al. (2012); Aguiar-Conraria et al. (2012)). Among the many recent economic applications using the continuous wavelet transform are, for example, Krüger (2021) with an evaluation of leading business cycle indicators for the German economy, Krüger and Neugart (2020) and Aguiar-Conraria et al. (2020) who examine the stability of Okun's law across time and frequency, and Verona (2016) and Flor and Klarl (2017) with financial applications.

There is also a growing literature where the continuous wavelet transform is applied in energy economics. For example, Papaioannou et al. (2015) examine the market integration between Italian and Greek electricity markets via bivariate wavelet analysis of their wholesale electricity prices. Sousa et al. (2014) apply bi- and multivariate wavelet tools to examine the link between EU ETS carbon prices and front-month future prices of electricity, gas, and coal. A key finding by Sousa et al. (2014) is that carbon prices and electricity prices are highly correlated at low frequencies, and remain so after controlling for the effect of the other variables, while the correlation between carbon price and gas or coal prices is significantly reduced after removing the influence of the other variables. Aguiar-Conraria et al. (2018) conduct a similar analysis for the California carbon market and examine the interaction of carbon prices with gasoline and electricity prices. Their key results from computing partial coherences can be summarized as follows. Gasoline and carbon prices have significant partial coherence at cycles with lengths around a year and half a year, where the two prices move in an out-of-phase relationship. For electricity and carbon prices they find that there is significant partial coherence at the same frequencies but over comparatively shorter portions of the sample period. At the half-year frequency the coherence and phase difference plots overall show an unstable relation between electricity prices and carbon prices. In contrast, at the yearly frequency the coherence is significant between mid 2015 and late 2016 (with a sample period covering 2014 - late 2017) where also the phase difference shows a stable in-phase relation.

Finally, Aguiar-Conraria et al. (2021) apply multivariate wavelet analysis to European, Japanese, and U.S. gas prices to examine the gas market integration. They first find from computing bivariate coherences the strongest connection between European and Japanese gas prices. Computing partial coherences then reveals that a large portion of this is actually driven by the Brent oil price, which itself is highly correlated with European and Japanese gas prices but less so with the U.S. gas price. When the influence of the Brent price is removed, significant coherence remains only at the longest cycles (6 - 9 years), and this result is alike for all three gas price pairs. This result from Aguiar-Conraria et al. (2021) exemplifies how bivariate coherence can give a "clouded" picture of the actual relations and illustrates the benefit of using partial coherences.

5.3 Wavelet and Cross-Wavelet Analysis

In this section, I explain the wavelet analysis tools that are used in this paper. Tools for uni- and bivariate time series are explained in sect. 5.3.1, while tools for multivariate time series are covered in sect. 5.3.2.

Wavelet analysis applied to time series can be regarded as a recent refinement of spectral analysis (see, e.g., Jenkins and Watts (1968); Koopmans (1974); Percival and Walden

(2020) for book-length treatments of the spectral analysis of time series). It can be seen as a refinement due to the fact that the wavelet transform allows to examine how the spectral properties of a time series change over time, thus providing resolution in both time and frequency domain. An earlier technique for providing time-frequency resolution is the so-called windowed Fourier transform, but the wavelet transform is advantageous. This is because the windowed Fourier transform uses windows of same width regardless of the frequency that is to be resolved, while the wavelet transform automatically compresses (stretches) the width of the wavelet when higher (lower) frequencies are to be resolved (see, e.g., Daubechies (1992), pp. 3f., for an illustration).

5.3.1 Uni- and Bivariate Wavelet Tools

The *continuous wavelet transform* (CWT) of a single time series $\{x_t\}$ in discrete time with $t = 1, \ldots, T$ represents the time series $\{x_t\}$ as a bivariate function in time and frequency space. The first step of the wavelet transform is the selection of a "mother wavelet", a continuous function $\psi(t)$, from which a family of "daughter wavelets" can be derived by $\psi_{s,\tau}(t) = |s|^{-1/2}\psi\big(\frac{t-\tau}{s}\big)$. Here, the parameter $\tau \in \mathbb{R}$ controls the time translation and $s \in \mathbb{R}$ controls the scale (inverse of frequency) where $|s| > 1$ stretches and $|s| < 1$ compresses the wavelet.

Several choices for $\psi(t)$ exist, which can be real-valued or complex-valued. In this paper, the analysis is carried out using the R-package 'WaveletComp' (version 1.1, see Rösch and Schmidbauer (2018)), which implements all wavelet transforms with the *Morlet mother wavelet*. The Morlet mother wavelet is the complex-valued function $\psi(t) = \pi^{-1/4}\mathrm{e}^{\mathrm{i}\omega_0 t}\mathrm{e}^{-t^2/2}$, where $\mathrm{i} = \sqrt{-1}$ denotes the imaginary number and ω_0 is a constant usually set to $\omega_0 = 6$ (Aguiar-Conraria and Soares (2014), p. 352).[74] First introduced by Grossmann and Morlet (1984), the Morlet wavelet is widely used due to its optimal resolution compromise between the time- and frequency-domain (Aguiar-Conraria and Soares (2014), p. 352). Letting * denote the complex conjugation, the CWT of the time series $\{x_t\}$ is defined formally as its convolution with the function $|s|^{-1/2}\psi^*\big(\frac{\cdot}{-s}\big)$, i.e.,

$$W_x(\tau, s) = \sum_{t=1}^{T} x_t \frac{1}{\sqrt{|s|}}\psi^*\Big(\frac{t-\tau}{s}\Big), \qquad \tau, s \in \mathbb{R},\ s \neq 0. \tag{5.1}$$

The WaveletComp package implements the CWT not by directly evaluating the convolution (5.1), but instead by multiplication in Fourier space and using Fast Fourier Transform (FFT) algorithms (see Torrence and Compo (1998) for details).

For complex-valued wavelets as the Morlet wavelet, the CWT is generally complex-valued, thus containing modulus and phase coordinates. In accordance with the terminology from spectral analysis, the *wavelet power spectrum* is defined as the squared modulus (divided by s for bias correction),[75] $P_x(\tau, s) = \frac{1}{s}|W_x(\tau, s)|^2$, and the *phase angle* is defined as

[74] To meet the formal admissibility condition for wavelets, the Morlet wavelet is more correctly defined as $\psi(t) = \pi^{-1/4}(\mathrm{e}^{\mathrm{i}\omega_0 t} - \mathrm{e}^{-\omega_0^2/2})\mathrm{e}^{-t^2/2}$ (Aguiar-Conraria and Soares (2014), p. 365), but for $\omega_0 = 6$ the correction term is so small that it can be ignored for numerical purposes.

[75] Division by s is implemented in the WaveletComp package as it reduces bias in the wavelet power spectrum, see Liu et al. (2007)

$\phi_x(\tau, s) = \arctan\left(\frac{\text{Im}(W_x(\tau,s))}{\text{Re}(W_x(\tau,s))}\right) \in (-\pi, \pi]$.[76] The wavelet power spectrum has the interpretation of an estimated variance as a function of time and frequency (inverse of scale), analogous to a Fourier power spectrum estimate which is interpretable as an estimated variance in the frequency domain only.

With more than a single time series, interest lies in applying the wavelet methodology to analyze how the covariance and correlation between the time series at different frequencies evolve across time. For two time series $\{x_t\}$ and $\{y_t\}$, the *cross-wavelet transform* (XWT) is defined in terms of the CWTs of the two time series, $W_x(\tau, s)$ and $W_y(\tau, s)$, simply as $W_{xy}(\tau, s) = \frac{1}{s} W_x(\tau, s) W_y^*(\tau, s)$.[77] The *cross-wavelet power spectrum*, given by the modulus of the XWT, is interpretable as an estimated covariance (or rather its absolute value) as resolved in frequency and time domain. This is analogous to the modulus of the sample cross spectrum in spectral analysis which provides a measure of covariation in the frequency domain only.

The measure of correlation analogous to the complex coherency from spectral analysis is the *complex wavelet coherency,* which is defined as

$$\varrho_{xy}(\tau, s) = \frac{S(W_{xy}(\tau, s))}{\sqrt{S(P_x(\tau, s)) \cdot S(P_y(\tau, s))}}, \tag{5.2}$$

where $S(\cdot)$ denotes the smoothing operation across scale and time.[78] As in the spectral analysis case, smoothing is necessary since otherwise ϱ_{xy} would have modulus of exactly one for all times and scales. Due to smoothing ϱ_{xy} is indeed a complex-valued correlation coefficient since it has modulus in $[0, 1]$, as follows from the Cauchy-Schwarz inequality.

As in Rösch and Schmidbauer (2018, p. 7), I will refer to the squared modulus of the complex wavelet coherency as the *wavelet coherence*, $R^2_{xy}(\tau, s) = |\varrho_{xy}(\tau, s)|^2$. This the key measure of "strength" of the co-movement between two time series after dissection into frequency- and time-domain. It is formally analogous to the coefficient of determination from statistics, as it represents information about the strength but not the sign of the correlation.

As a measure of the sign of the correlation, using a complex-valued wavelet allows to quantify the lag-lead relationship between two time series. In particular, this can uncover whether the co-movement relation at a specified frequency band is in-phase (positive correlation) or out-of-phase (negative correlation), and whether this relationship is stable or unstable over time. In this paper, the *phase difference* between series x_t and y_t is computed as the phase coordinate of the complex wavelet coherency,[79]

$$\phi_{xy}(\tau, s) = \arctan\left(\frac{\text{Im}(\varrho_{xy}(\tau, s))}{\text{Re}(\varrho_{xy}(\tau, s))}\right) \in (-\pi, \pi]. \tag{5.3}$$

[76] The convention is that the angle lies in $(0, \pi)$ (or $(-\pi, 0)$, respectively), if the imaginary part of the complex number is positive (or negative, respectively).

[77] Division by s is implemented in the WaveletComp package also for the bivariate case, also for bias-correction, see Veleda et al. (2012).

[78] Intuitively, the smoothing operation can be thought of as producing a weighted average from a finite number of equally-spaced points located within a certain window around the point (τ, s).

[79] Some papers instead define the phase difference as the phase coordinate of the XWT, which can differ somewhat from my definition since the coherency is affected by smoothing. I use the definition in (5.3) because only this can be carried over to the multivariate case (see remark 2.2 in Aguiar-Conraria et al. (2018)).

If the phase difference is exactly zero then both series move in perfect lockstep, while for a phase difference of π or $-\pi$ both series move perfectly anti-cyclical. Making the abbreviation $\phi_{xy} \equiv \phi_{xy}(\tau, s)$, we can summarize the lag-lead relationship as follows (Aguiar-Conraria and Soares (2014), p. 356): both series move in-phase with x_t leading if $\phi_{xy} \in (0, \pi/2)$, and they move in-phase with y_t leading if $\phi_{xy} \in (-\pi/2, 0)$. An out-of-phase relationship is present if $|\phi_{xy}| > \pi/2$, where x_t is leading if $\phi_{xy} \in (-\pi, -\pi/2)$, and y_t is leading if $\phi_{xy} \in (\pi/2, \pi)$.

Statistical significance of the estimated coherences and wavelet power spectra can be assessed via Monte Carlo simulation of surrogate series, as is implemented with several options in the R-package WaveletComp (Rösch and Schmidbauer (2018), p. 8). For the wavelet power spectra of individual time series I select the option 'AR(1)', where the surrogate series are red noise processes fitted to the original time series. To judge the significance of the coherences WaveletComp offers the option 'Fourier randomization', where two independent surrogate series that have the same Fourier power spectrum as the two original series are simulated.

The significance tests for the bivariate and multivariate applications carried out in this paper are not based on the simulations done using the WaveletComp package. Since the methods in WaveletComp do not save the simulated surrogate series, I carry out my own implementation of 'Fourier randomization' because this is necessary to assess the significance when applying the multivariate wavelets tools that will be explained in the next subsection.

My own implementation of 'Fourier randomization' is based on the 'circulant embedding' method as described in Percival and Walden (2020) (pp. 601-611), which can be implemented efficiently via FFT algorithms. In the empirical applications I first of all take first differences for all of the time series. Afterwards I remove whatever trend remains using the R-command 'loess' (with span=0.75) and also standardize the time series, since this is what is done internally by the methods in WaveletComp (Rösch and Schmidbauer (2018), p. 5). For the resulting time series, the simulation method then generates surrogate series of the same length, which are distributed as zero mean Gaussian stationary processes whose Fourier power spectrum is identical to the estimated Fourier power spectrum of the original series.[80] For this, I estimate the Fourier power spectrum of the original series by the periodogram (e.g., Percival and Walden (2020), p. 170). Since the different series are generated independently, this allows to derive the Monte Carlo distribution of the wavelet coherence (as a function of (τ, s)) under the null hypothesis of no coherence between any of the series, while permitting auto-correlation (i.e. nontrivial univariate Fourier power spectra) in each series.

In all simulations I use 1000 repetitions. In order to keep the simulations of the coherencies computationally tractable, I make the following approximation: after computing the wavelet coherency between a pair of surrogate series, I only use the values from each 20^{th} point (τ) in time-space and each 5^{th} point (s) in frequency-space. I then approximate the wavelet coherencies for all points in-between by linear interpolation. This is unlikely to create large error since the CWT (eq. 5.1) is continuously differentiable. Indeed, by comparing the areas of statistical significance (in the bivariate coherence plots) which result from my own simulations with those computed using the 'Fourier.rand' option of

[80] By implication, the process also has an auto-covariance function that is identical to the estimated auto-covariance function of the original series.

the WaveletComp package, I observe that the areas are nearly identical, which serves as a check on the accuracy of my simulation procedure.

5.3.2 Multivariate Wavelet Tools

When one is concerned with more than two time series there can be interactions between any of them, thus there is a need to account for these interactions in the wavelet analysis. Luckily, since complex wavelet coherency is formally entirely analogous to the correlation coefficient, as is the case for Fourier coherency in spectral analysis, the techniques from linear regression theory and from correlation analysis based on the multivariate normal distribution can be applied here (Aguiar-Conraria and Soares (2014), p. 357). Thus, partial coherency and multiple coherence can be obtained in the same way as partial and multiple correlation coefficients can be calculated from a correlation matrix (see, e.g., Kendall and Stuart (1979), ch. 27; Jenkins and Watts (1968), pp. 476ff.).

In the following, let there be multiple time series where the complex wavelet coherency between any pair of series exists. To simplify the notation I drop the dependence on (τ, s) so that coherency, coherence, and phase difference for the pair (i, j), $i \neq j$ are denoted $\varrho_{ij} \equiv \varrho_{ij}(\tau, s)$, $R^2_{ij} \equiv R^2_{ij}(\tau, s)$, and $\phi_{ij} \equiv \phi_{ij}(\tau, s)$, respectively. For example, if one wants to calculate the coherency between series 1 and 2 after controlling for the influence of series 3, the *partial complex wavelet coherency* denoted by $\varrho_{12.3}$, one can use the following formula (Aguiar-Conraria and Soares (2014), p. 358)

$$\varrho_{12.3} = \frac{\varrho_{12} - \varrho_{13}\varrho^*_{23}}{\sqrt{(1 - R^2_{13})(1 - R^2_{23})}}. \tag{5.4}$$

For the general case with p series, let $\boldsymbol{A}$ denote the $p \times p$-matrix containing the coherencies,

$$\boldsymbol{A} = \begin{pmatrix} 1 & \varrho_{12} & \cdots & \varrho_{1p} \\ \varrho^*_{12} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \varrho^*_{1p} & \cdots & \cdots & 1 \end{pmatrix}.$$

Note that this is formally a complex-valued correlation matrix and thus Hermitian (and positive definite). Partial coherencies can then be expressed in terms of the cofactors of $\boldsymbol{A}$. Let $\det(\cdot)$ denote the determinant, and let $\boldsymbol{A}_{ij}$ be the submatrix of $\boldsymbol{A}$ obtained by deleting the i^{th} row and the j^{th} column, then the cofactor to the $(i, j)^{\text{th}}$ entry in $\boldsymbol{A}$ is defined as $(-1)^{(i+j)}\det(\boldsymbol{A}_{ij})$. I also use $\boldsymbol{q}_j$ as an abbreviation for all series except 1 and j, i.e. $\boldsymbol{q}_j = \{2, \ldots, p\} \backslash \{j\}$. Then, one can compute the *partial complex wavelet coherency* between series 1 and j, controlling for the influence of all other series, by (Aguiar-Conraria and Soares (2014), p. 369)[81]

$$\varrho_{1j.\boldsymbol{q}_j} = -\frac{(-1)^{(j+1)}\det(\boldsymbol{A}_{j1})}{\sqrt{\det(\boldsymbol{A}_{11}) \cdot \det(\boldsymbol{A}_{jj})}}. \tag{5.5}$$

[81] The partial coherency $\varrho_{1j.\boldsymbol{q}_j}$ involves the cofactor to the $(j, 1)^{\text{th}}$ entry in $\boldsymbol{A}$ (i.e., the order is swapped) because this cofactor appears as the $(1, j)^{\text{th}}$ entry in the inverse $\boldsymbol{A}^{-1}$(up to the constant $1/\det(\boldsymbol{A})$). See Kendall and Stuart (1979), ch. 27 for a derivation for the case of real-valued partial correlations. Since $\boldsymbol{A}$ is Hermitian it follows that $\det(\boldsymbol{A}_{1j}) = \det(\boldsymbol{A}_{j1})^*$, and thus from (5.5) that $\varrho_{j1.\boldsymbol{q}_j} = \varrho^*_{1j.\boldsymbol{q}_j}$, which shows that the partial coherencies are conjugate symmetric just like the bivariate coherencies, resulting again in a Hermitian matrix of partial coherencies.

Table 5.1: Time series used in the wavelet analysis

Abbreviation	Time series	Unit
ELP	Wholesale electricity price	ct/kWh
TOT	Total electricity load	GWh
VRE	Variable renewable electricity generation[a]	GWh
RES	Residual load[b]	GWh
COP	Rotterdam coal futures price	€/ton
NGP	Natural gas spot price	€/mill BTU

[a]Defined as electricity generation from onshore wind, offshore wind, and solar photovoltaic.
[b]Defined as TOT minus VRE.

It is easy to see that (5.4) is a special case of (5.5) for $j = 2$ and $\boldsymbol{q}_j = \{3\}$.

As above, the *partial wavelet coherence* and *partial phase difference* are derived from this by $R^2_{1j.\boldsymbol{q}_j} = |\varrho_{1j.\boldsymbol{q}_j}|^2$ and $\phi_{1j.\boldsymbol{q}_j} = \arctan\left(\frac{\mathrm{Im}(\varrho_{1j.\boldsymbol{q}_j})}{\mathrm{Re}(\varrho_{1j.\boldsymbol{q}_j})}\right) \in (-\pi, \pi]$. Turning now to the multiple wavelet coherence, this measure is defined analogously to the multiple R^2 in linear regression. Denoting the *multiple wavelet coherence* between series 1 and all the other series by $R^2_{1(2\ldots p)}$, this can be computed from $\boldsymbol{A}$ by (Aguiar-Conraria and Soares (2014), p. 369)

$$R^2_{1(2\ldots p)} = 1 - \frac{\det(\boldsymbol{A})}{\det(\boldsymbol{A}_{11})}. \tag{5.6}$$

Alternatively, it can be calculated from partial wavelet coherences by the recursive formula (Aguiar-Conraria and Soares (2014), p. 370)

$$R^2_{1(2\ldots p)} = 1 - (1 - R^2_{12})(1 - R^2_{13.2}) \ldots (1 - R^2_{1p.(2\ldots(p-1))}). \tag{5.7}$$

Finally, I also want to compute the multiple wavelet coherence between series 1 and series $2, ..., (p-1)$, after controlling for the influence of series p. This can be done by first computing the partial coherencies $\rho_{ij.p}$ for all (i, j) with $1 \leq i < j \leq p-1$, and then plugging the resulting Hermitian correlation matrix into (5.6).

5.4 Time Series Data

In this section, I describe the time series data that are used in the subsequent wavelet analysis. The time series are listed and abbreviated in Table 5.1. The first four are electricity market data which are published online by the German Federal Network Agency (Bundesnetzagentur).[82]

These are daily time series on the wholesale electricity price (which is fixed at the European Energy Exchange (EEX) one day ahead at 12:00), the total electricity load, variable renewable electricity generation, and the residual load. I obtain full data coverage for all days from the 5$^{\text{th}}$ of January 2015 until the 29$^{\text{th}}$ of December 2023, so that each time series has a length of 3281 observations. The total electricity load refers to the aggregate

[82]The data are available from 2015 onwards and can be downloaded from https://www.smard.de/home/downloadcenter/download-marktdaten/

Table 5.2: Yearly total electricity load and electricity generation shares for Germany

	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total electricity load [TWh]	495	503	506	509	497	485	505	483	455
Thereof, ...									
hydro + biomass + other RE [%]	10	12	11	11	12	12	11	11	12
VRE (all wind + solar) [%]	25	25	29	31	35	38	34	40	45
hard coal + lignite [%]	43	42	39	39	30	24	30	34	26
natural gas [%]	3	5	5	8	11	12	10	8	11
nuclear [%]	17	16	14	14	14	13	13	7	1
other non-renew. [%]	11	9	2	2	3	3	3	2	3
net imports [%]	-4	-4	3	-3	-2	-1	4	1	5
net PHS generation[a] [%]	-4	-4	-4	-3	-3	0	-4	-3	-3

[a]PHS=pumped hydro storage.

Source: German Federal Network Agency (Bundesnetzagentur), https://www.smard.de/

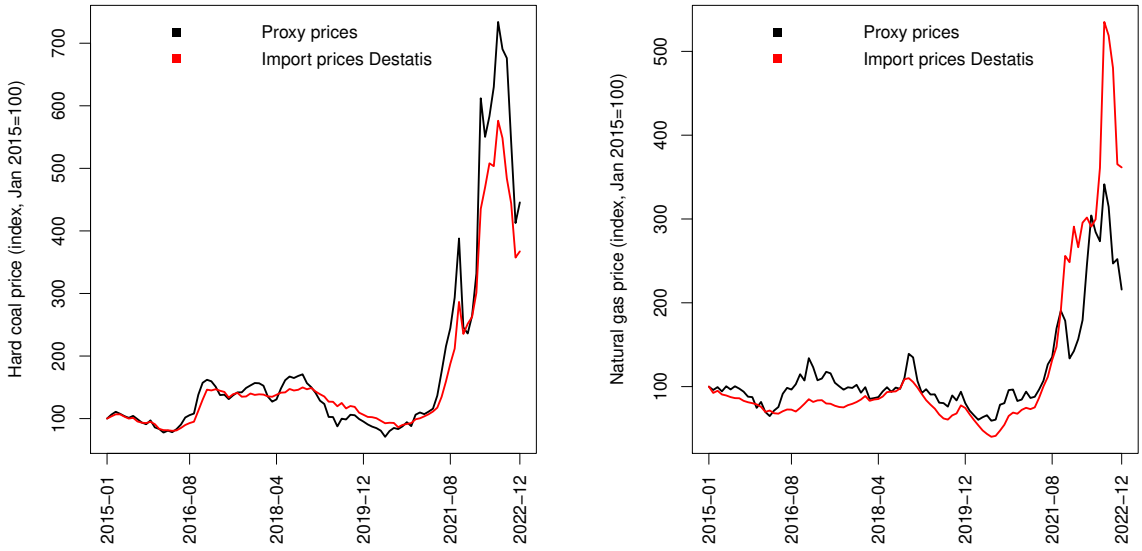
quantity of electricity that is taken from the Germany power grid by all power consumers situated in Germany, thus it reflects total German electricity consumption or demand. By definition it also equals total generation from all electricity generation units (including net generation from pumped hydro-storage) plus the net electricity imports from neighboring countries. Variable renewable electricity (VRE) generation includes onshore wind, offshore wind, and solar photovoltaic. The residual load is defined as the total load minus VRE generation. Therefore, it includes the generation from biomass, hydropower, net generation from pumped hydro-storage, nuclear power plants, the major fossil power plants (lignite, hard coal, and natural gas), other fossil or non-renewable power plants like diesel-based generators or waste incineration, and the net electricity imports from neighboring countries. Table 5.2 shows the total electricity load for each year and the shares of the different generation technologies and net imports in the total load.

I complement this dataset with daily time series that are proxies for the import prices of natural gas and hard coal purchased from the international market. This is to reflect these procurement prices as the major items of variable cost for German fossil power generation plants. For the proxy series for natural gas import prices I obtain daily natural gas spot prices published online by the German exchange in Frankfurt.[83] For hard coal I could not find any data source which publishes spot prices. Therefore, I use prices of Rotterdam coal futures traded on the Intercontinental Exchange (ICE). The series is obtained from https://www.tradingview.com/. On the major commodity exchanges like the ICE, one can trade distinct future contracts with different maturities. The contracts are differentiated by expiration month and mature at the end of the expiration month. The website https://www.tradingview.com/ offers price information on "continuous future contracts", which are artificial financial instruments that concatenate different future contracts upon maturity. In particular, I download the time series that has the abbreviation "ATW1!" on the website. This is a so-called "front-month continuous contract" that automatically rolls over to the future contract with the nearest expiration month once the current contract expires.[84] In this sense, the series should be the closest available proxy to a spot price series.

[83]This is downloaded from https://www.boerse-frankfurt.de/rohstoff/erdgaspreis/

[84]For further explanation see https://www.tradingview.com/support/solutions/43000483493-what-are-1-and-2-continuous-futures-contracts/

Figure 5.1: Comparison of proxy prices to monthly import prices



Both price series are denominated in US$, so I apply daily exchange rates downloaded from the ECB Data Portal to convert them into €.[85] Because the coal price, gas price, and exchange rate data refer to exchange-traded contracts, no data exists for weekends and national holidays. Both the coal and gas price data contain an opening and a closing price for each trading day from which I construct artificial prices for weekends and national holidays by linear interpolation between the last available closing and the next available opening price. For each trading day I use the average of opening and closing price to arrive at a single series. The result of this procedure are two price series (COP and NGP) denominated in € with full data coverage for each day between 5th Jan 2015 and 29th Dec 2023.

To assess how well the proxy price series reflect the actual procurement prices of German fossil generation plants, I also obtain monthly data from the Federal Statistical Office of Germany (Destatis) on import price indices for hard coal and natural gas up until the end of 2022.[86] Figure 5.1 compares the monthly series obtained from Destatis to the monthly series derived from the proxy price series by calculating monthly averages. The figure shows that monthly import prices for hard coal are very well represented by the prices of Rotterdam coal futures, at least until the end of August 2021 after which proxy prices are mostly above import prices. The correlation between both series is 0.99. Import prices for natural gas are somewhat less well represented by natural gas spot prices with a correlation of 0.93. Especially noticeable are again the differences after August 2021 where now proxy prices are substantially below import prices. Overall, I conclude from Figure 5.1 that both series are reasonable proxies for the long-run behavior of the procurement prices. However, to the extent that procurement prices are fixed in the short-run via bilateral contracts, the short-term movements in the proxy prices contain volatility that is unlikely to be present in the actual procurement prices, and is unlikely to be passed on to electricity prices by the German utilities. As a consequence for the subsequent wavelet analysis, I expect to find high values for the coherence between the proxy price series and

[85] The series is accessible at https://data.ecb.europa.eu/data/datasets/EXR/EXR.D.USD.EUR.SP00.A/

[86] This is downloaded from https://www.destatis.de/EN/Themes/Economy/Prices/Publications/Downloads-Energy-Price-Trends/energy-price-trends-pdf-5619002.html

electricity prices only for the longer periods (e.g., above one year).

5.5 Empirical Results

In this section, I report the results from conducting a wavelet analysis with the time series listed in Table 5.1. In sect. 5.5.1, I first analyze the results of applying univariate wavelet tools to each time series. I also discuss similarities and correlations between the time series that can be seen either from the series directly or from their (wavelet) power spectra. In sect. 5.5.2, I discuss the results of applying bivariate wavelet tools (i.e., coherence and phase difference analysis) to the electricity price paired with any of the other series (i.e., the explanatory variables). Finally, sect. 5.5.3 examines how the results from the bivariate analysis can be refined by applying the multivariate wavelet tools introduced in sect. 5.3.2. This is important because there can be interdependencies among any of the explanatory variables, which could cause the bivariate analysis to give a clouded picture of the actual relevance of the different variables. Also, I want to quantify the *joint* explanatory power that the input prices and the quantity series have for explaining the electricity price, which due to interdependencies among the variables cannot be inferred from the *individual* explanatory powers.

5.5.1 Univariate Wavelet Analysis

Figures 5.2 and 5.4 depict the results of the univariate wavelet analyses for the time series listed in Table 5.1. To assure stationarity of all time series, the series ELP, RES, VRE, and TOT are taken as first differences, and the series NGP and COP are log-transformed and then taken as first differences.[87] The left panels in each figure plots the time series as handed over to the R-command for computing the wavelet transform. The middle panels of the figures show the colored plots of the wavelet power spectra, where the correspondence between colors and numerical values can be found on right-hand side of the plot. White lines in the plot indicate the ridges (i.e., local maxima) of the wavelet power spectrum, and the shaded, purple-colored area is the so-called cone of influence. This is the area of time-period space where results can only be interpreted with caution due to the edge effects caused by the finite length of the time series (e.g., Torrence and Compo (1998)). Also drawn in the middle panels are black contour lines that enclose areas of statistical significance at the 10% level. As explained in sect. 5.3, statistical significance is assessed from comparing the wavelet power spectra to those of 1000 simulated red noise (i.e., AR(1)) processes. The right panels of the figures show the time-averaged wavelet power spectrum, which I refer to for briefness simply as the power spectrum. This constitutes an estimate of the spectral density of the time series which can be interpreted just as "classical" spectral density estimates based on the Fourier transform. Note that in contrast to the two other panels, the abscissa here shows the values of the power spectrum instead of calendar dates. Statistical significance is depicted by the colored dots that are superimposed onto the graph, for which blue and red dots indicate significance at the 10%- and 5%-level, respectively.

[87] Table D1 in Appendix D reports the results of unit-root tests carried out with the R-package 'urca'. The results confirm that all series are stationary after taking first differences. Since wavelet analysis can also deal with nonstationary time series, I also show the analogous figures of the univariate wavelet analysis with time series in levels in Figures D1-D2 in Appendix D.

Figure 5.2: Univariate wavelet results

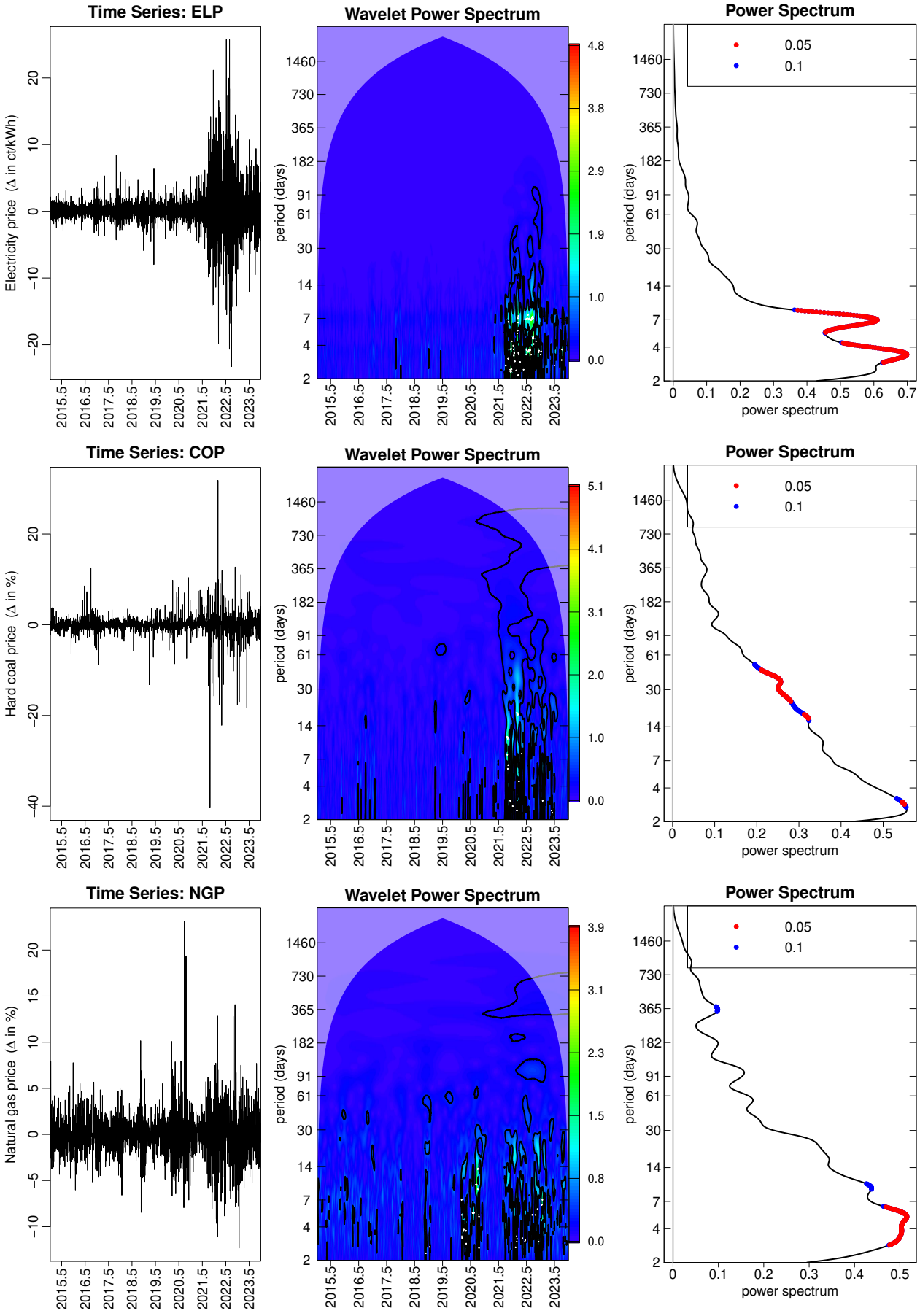
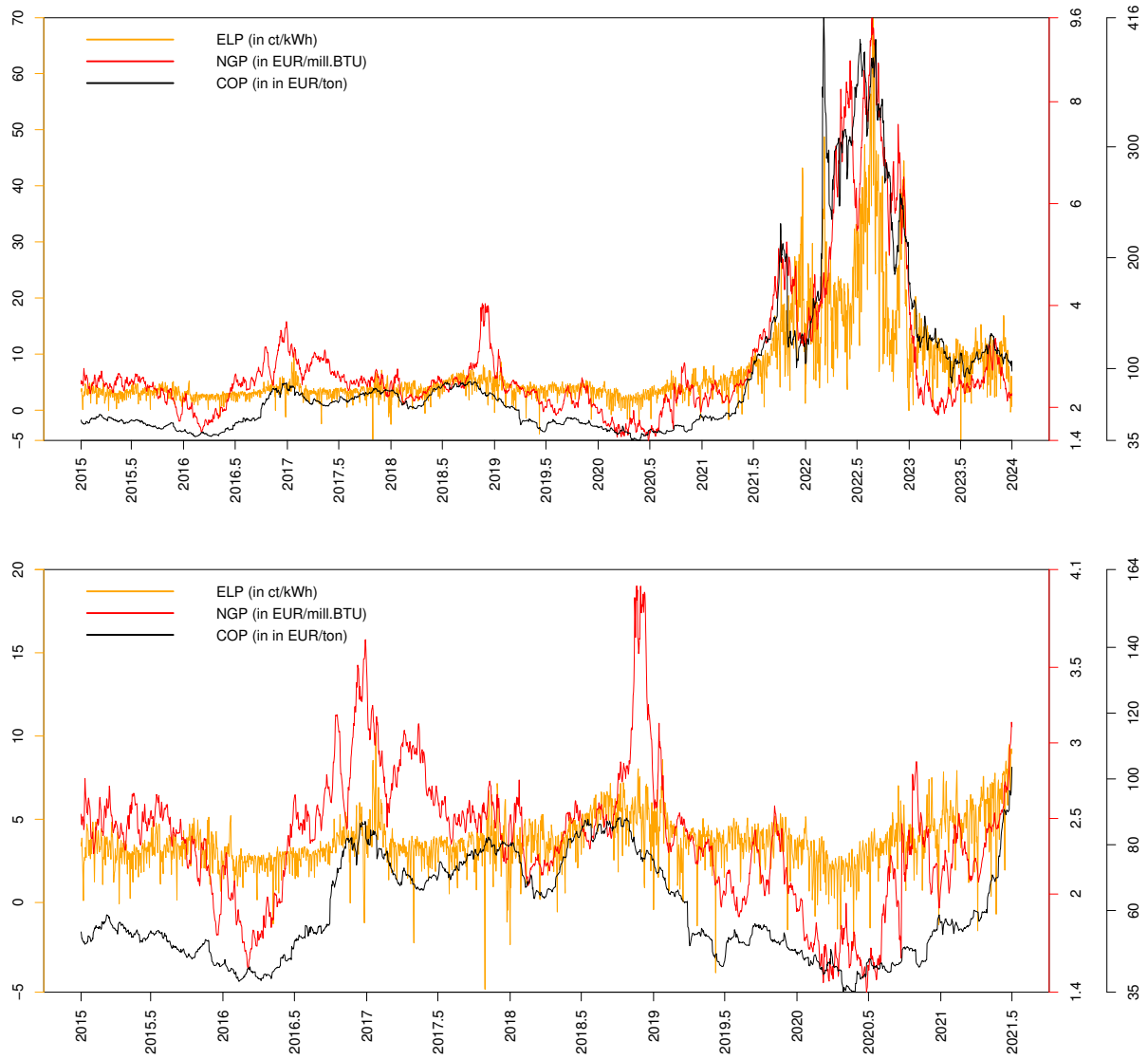
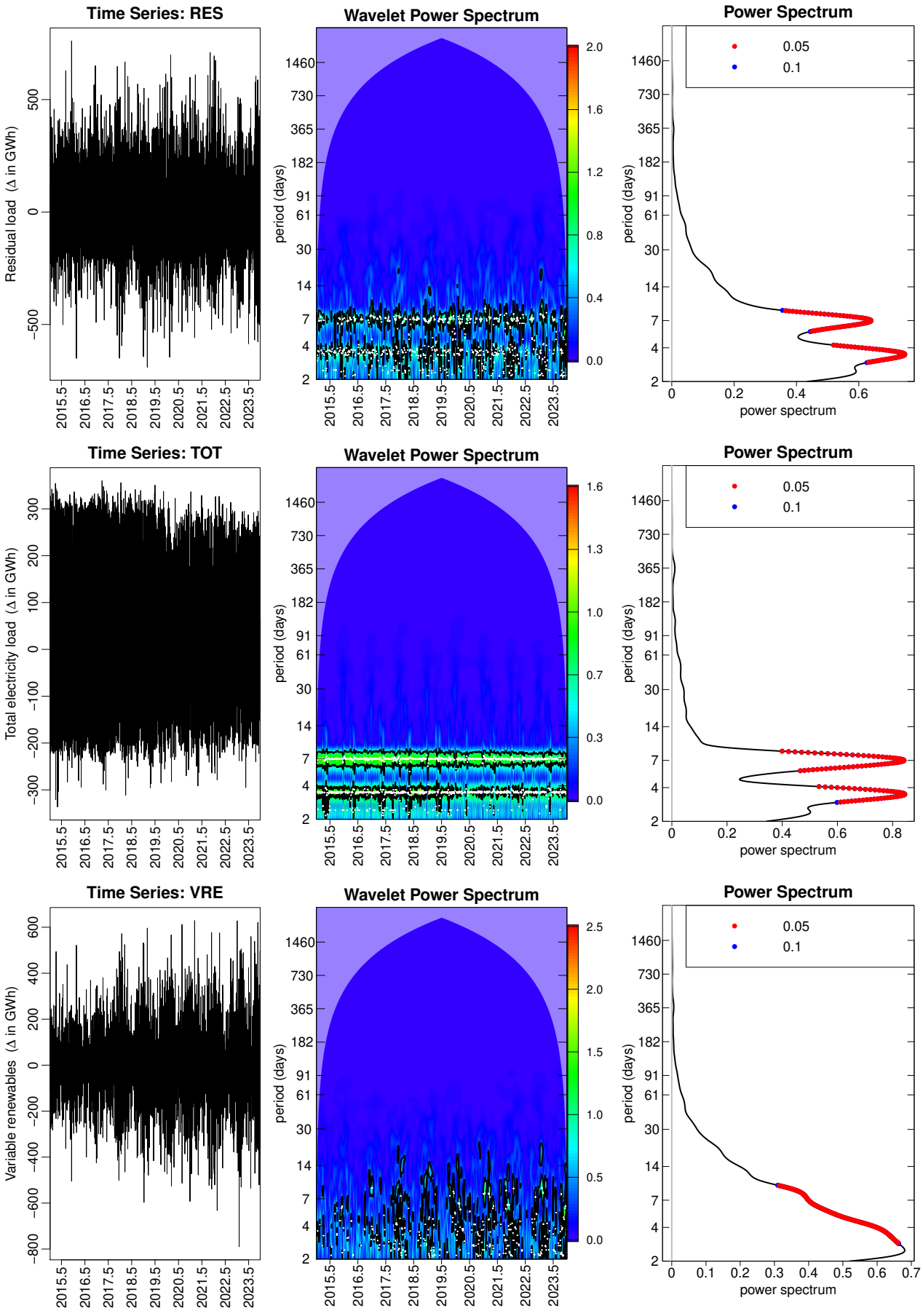


Figure 5.3: Wholesale electricity, natural gas, and hard coal prices in levels



Note: the lower panel "zooms in" on Jan 2015 - July 2021, the time period before the price explosion. The colors of the axes match the colors of the time series. Units are given in the legend.

Figure 5.4: Univariate wavelet results (cont'd)



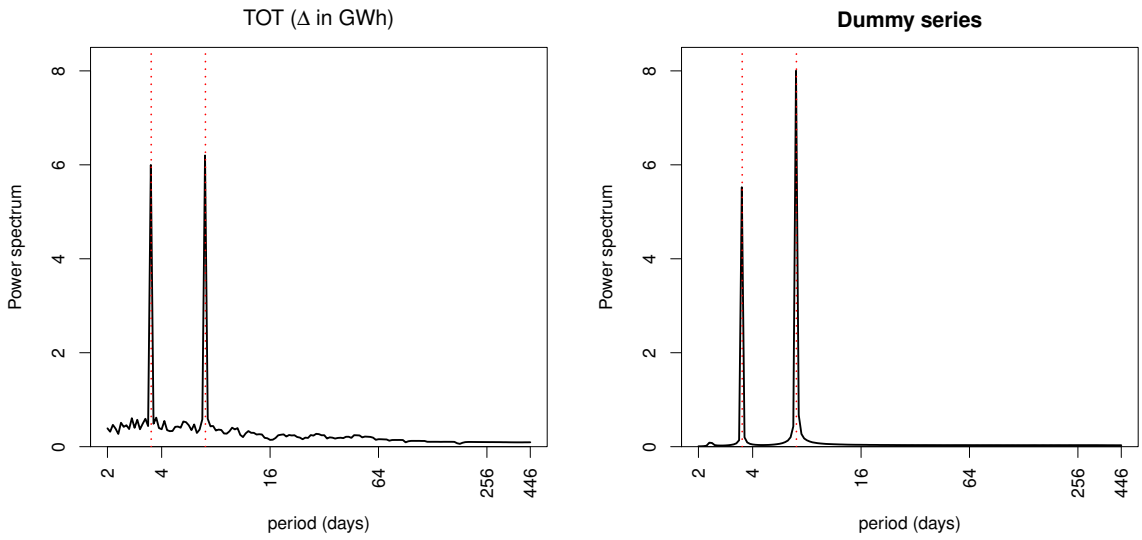
The top-left panel of Figure 5.2 shows that day-to-day electricity price changes have undergone a structural break in form of a large increase in volatility starting in mid 2021. Consequently, the top-middle panel of Figure 5.2 shows a visibly large and significant wavelet power spectrum at the shorter periods only from mid 2021 onwards. Considering the electricity price series expressed in levels, which is shown in Figure 5.3 as the orange-colored curve, reveals that there has been a strong price increase after mid 2021. The average wholesale electricity price in 2015-2020 was fairly stable around a value of 3.5 ct/kWh. Thereafter, it increased to 9.7 ct/kWh in 2021, to 23.5 ct/kWh in 2022, and returned back to 9.6 ct/kWh in 2023. In Figure 5.3 I also overlay the electricity price series with the natural gas and hard coal price series, shown in red and black, respectively. As can be seen immediately from the figure, the price explosion starting in 2021, accelerating in 2022, and the subsequent decline in 2023 is shared by all three series. Thus, an evident interpretation of Figure 5.3 is that the electricity price anomaly in 2021-2023 has been caused by the movements of gas and coal prices. This will be further examined using bi- and multivariate wavelet tools in sects. 5.5.2-5.5.3. That the increased price volatility on international energy commodity markets had already started before the Russian invasion of Ukraine (which began on 24$^{\text{th}}$ February 2022), namely in the second half of 2021, is also found by several papers examining the pandemic-era inflation, e.g. Ball et al. (2022), Blanchard and Bernanke (2023), and di Giovanni et al. (2023).

In Figure 5.2, the rows below continue with the results for day-to-day percentage changes in coal (middle row) and gas prices (bottom row). From the discussion above, Figure 5.3, and the left panels in Figure 5.2, it is now clear why also here the wavelet power spectra show the highest volatility during the era between mid 2021 and mid 2023. Gas prices also exhibit high volatility around mid 2020 with some exceptional day-to-day changes above +15%. Note from Figure 5.3 that this occurs after gas prices have reached the minimum price of the sample period in June 2020, so the increase results as a catch-up effect from a very low price level. The low gas and coal prices during the first half of 2020 are clearly related to the shock of the COVID-19 pandemic and the subsequent lock-down policies, which reduced demand and demand expectations for energy. For example, oil prices even briefly were negative during April 2020.[88] Turning next to the power spectra, the middle-right panel of Figure 5.2 shows a power spectrum for coal prices which is statistically larger than that of red noise chiefly for the 14-60 days period band. The bottom-right panel shows that for gas prices this is the case mainly for the 2-7 days period band, but also for the period of a year. A plausible explanation here is that the widespread use of gas for the heating of buildings in Germany induces important yearly periodicity into the gas price series. Overall, Figure 5.2 reveals that the power spectra of the two price series COP and NGP put higher weight on larger periods as compared to that of the electricity price series.

Turning now to Figure 5.4 with the electricity-quantity time series, and first to the middle row for the total load, one sees that day-to-day changes of the total load rarely exceed ±300 GWh (compared to an average daily total load for 2015-23 of 1353 GWh). In comparison, the residual load and VRE generation exhibit a larger range of day-to-day changes caused by the inherent intermittency of the variable renewables, which I will analyze in more detail in a paragraph below. Considering beforehand the (wavelet) power spectra of the residual and the total load, one sees at first that these are fairly stable over time and, secondly, that these show pronounced maxima at the 7- and 3.5-day periods.

[88] See https://www.eia.gov/todayinenergy/detail.php?id=46336.

Figure 5.5: Fourier power spectra of the total load (day-to-day changes) and of a 7-day periodic dummy series



Note: the dotted red lines are at 3.5 and 7 days. The dummy series is a 7-day periodic series with value 1 at days 1-5 and value -1 at days 6-7, which is repeated to give a series of 3820 observations.

It is evident that the residual load "inherits" this from the total load and that the 7-day periodicity in the latter is caused by the weekly cycle of electricity demand with higher demand on work days than on weekends.

What explanation can be found for the importance of the 3.5-day period? It turns out that this is largely an artifact due to the 7-day periodicity being composed of 5 work days followed by 2 days of weekend. In Figure 5.5, I compare the power spectra of day-to-day changes in the total load with that of a 7-day periodic dummy series that takes on the value 1 for 5 days and then equals -1 for 2 days. Both power spectra are computed as simple periodograms via the discrete Fourier transform, are then smoothed across frequency, and plotted on a log-scaled period axis. The figure reveals the clear-cut similarity between the two power spectra, as both are dominated by two peaks located precisely at 3.5 and 7 days.

Considering next the results for VRE generation depicted in the bottom row of Figure 5.4, one can see that day-to-day changes on many days exceed ±300 GWh. Besides, it is visible from the left and the middle panel that the volatility of VRE generation has gradually increased over the sample period. Turning to the power spectrum on the right one sees that, much in contrast to what is the case for the total load, the power spectrum declines gradually, almost linearly, for the periods between 3-30 days. This shows how the inherent intermittency caused by weather-dependency is reflected in a power spectrum with a broader range of periodicities. To examine the VRE time series and in particular its volatility in more detail, I show in Table 5.3 several statistics for 2015-23. In Table D2 in Appendix D, I also report the same statistics for the three components solar, onshore wind, offshore wind, and for comparison also for nuclear power. In the tables, the first row showing the installed generating capacity is based on data that is also available at https://www.smard.de. The rows below report the mean, the capacity factor, and the coefficient of variation for daily generation. Rows 1-3 of Table 5.3 show how the mean

Table 5.3: German VRE generation capacity, mean and variability of daily generation

	2015	2016	2017	2018	2019	2020	2021	2022	2023
Installed generating capacity[a] [GW]	79.6	87.6	95.7	102.0	106.7	112.2	118.2	124.8	135.5
Daily generation, mean [GWh]	340	338	404	432	476	504	463	523	567
Daily generation, capacity factor [%]	18	16	18	18	19	19	16	17	17
Daily generation, coef. of var. (CV) [%]	42	42	43	41	42	42	44	40	39
Daily generation, robust CV[b] [%]	28	25	30	26	27	29	29	28	27

[a]Reported are the means between two consecutive years as the original data refer to installed capacity on 1st of January.
[b]Defined as the interquartile range divided by the sum of the quartiles.
Source: German Federal Network Agency (Bundesnetzagentur), https://www.smard.de/

daily VRE generation rises alongside the installed generation capacity at a fairly stable proportion, with an average capacity factor over the sample period of 17.5%. The last two rows report the coefficient of variation (CV) and a robust version of this statistic (e.g., Arachchige et al. (2022)). Both coefficients lie within a range of 5 percentage points, which suggests also a fairly proportional relationship between volatility and mean of daily VRE generation.

To conclude this subsection, note that from the relation RES = TOT − VRE and the right panels in Figure 5.4, one can see how the power spectrum of the residual load is a "mixture" of the more narrow spectrum of the total load and the broader spectrum of VRE generation. Besides, comparing the power spectrum of the residual load with that of the electricity price in Figure 5.2 reveals that both look strikingly similar, and in fact from all pairs that involve the electricity price this is the pair with the largest overlap in the power spectra. Hence, I expect that this is the pair where the bivariate wavelet analysis shows the highest coherence over a broad range of periods. As discussed above in connection with Figure 5.3, I also expect to find high coherence between electricity prices and coal and gas prices at the longer periods.

5.5.2 Bivariate Wavelet Analysis

In this section, I separately examine the influence of each of the quantity series (RES, VRE, and TOT) and each of the price series (COP and NGP) on the electricity price via bivariate coherence and phase difference analysis.

Figures 5.6-5.7 depict on the left the pairwise wavelet coherences for several pairs of time series. The calculated wavelet coherence is depicted on a color scale ranging from blue (zero) to red (one). As above, the now light-shaded area is the cone of influence where the results can only be interpreted with caution due to edge effects, and the black contour lines enclose areas where the wavelet coherence is statistically significant at the 10% level. As explained in sect. 5.3.1, this is assessed from 1000 simulations of two independent surrogate series with a similar spectrum as the original two series. Inside the enclosed areas, the black arrows point in the direction of the phase difference between the two time series at the specific point in time-period space, as calculated from eq. (5.3). On the right, the figures show how the phase differences at certain period bands evolve over the sample period. For this, the phase differences are averaged over all periods within a certain interval (period band) that is always annotated as the label of the y-axis. To repeat the explanation from sect. 5.3, if the phase difference between x and y is between

Figure 5.6: Bivariate wavelet results

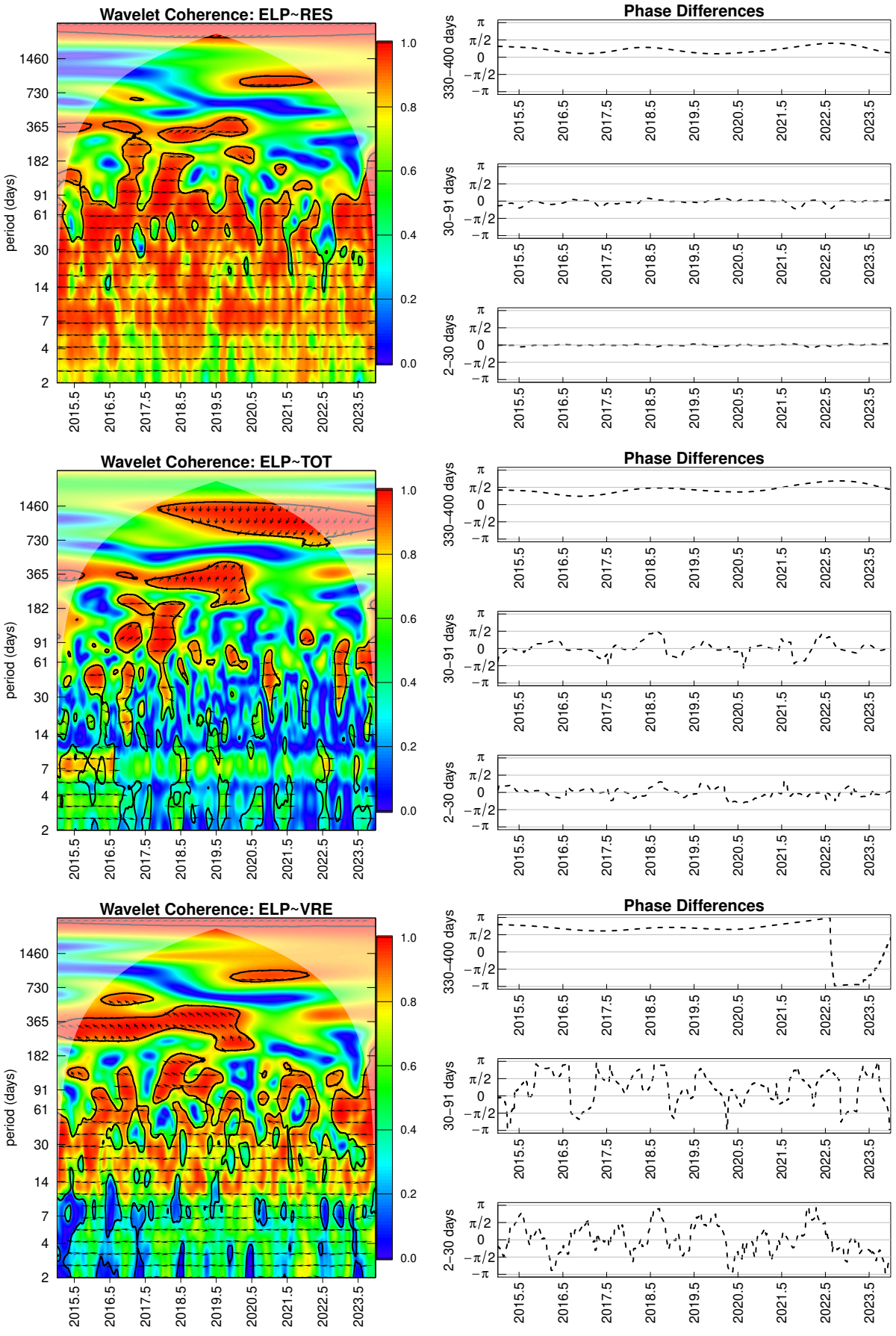
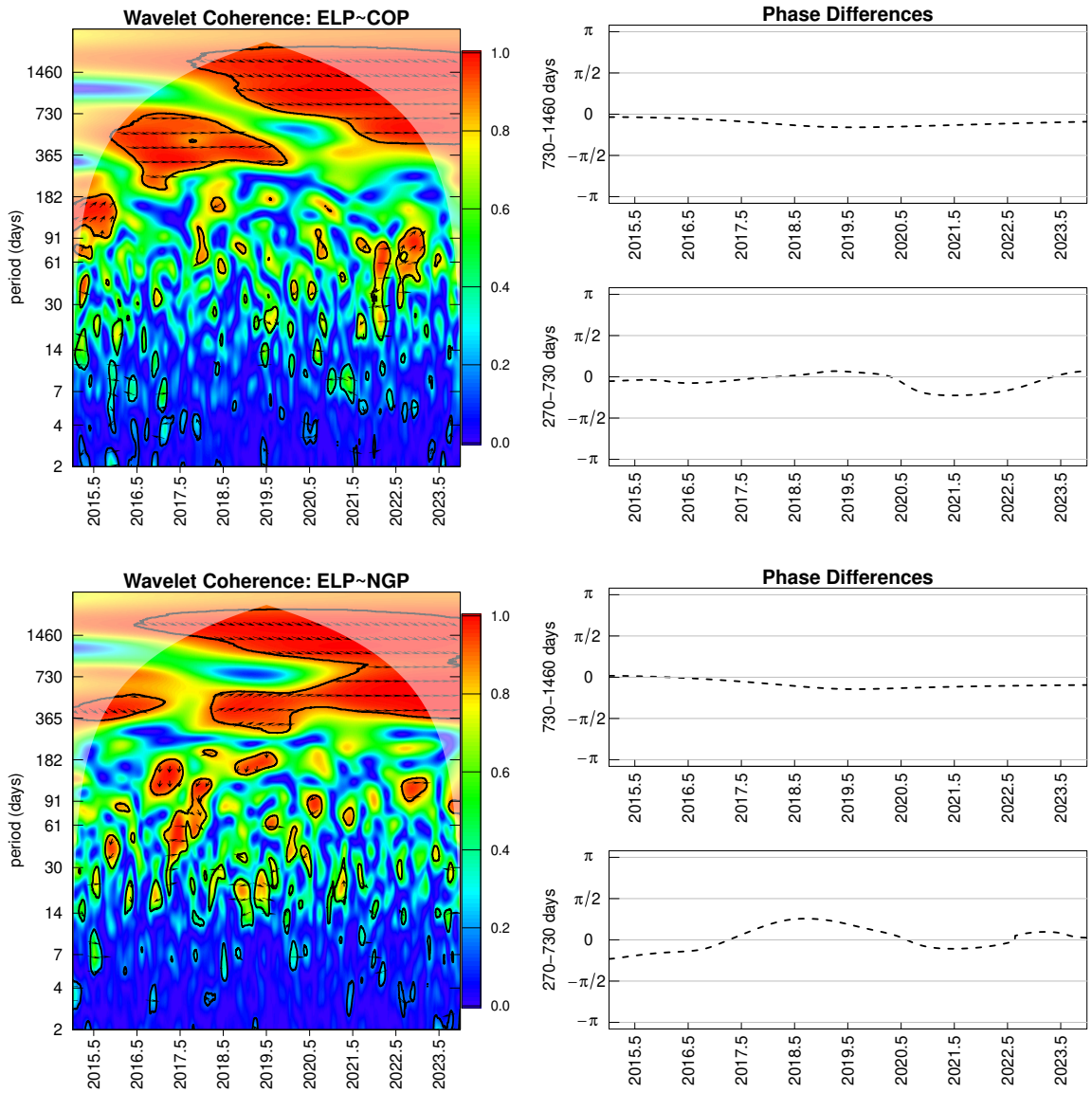


Figure 5.7: Bivariate wavelet results (cont'd)



0 and $\pi/2$ ($-\pi/2$ and 0) the series are said to move in-phase, that is they are positively correlated, and x (y) is leading. If the phase difference is between $\pi/2$ and π ($-\pi$ and $-\pi/2$) the series are said to move out-of-phase, that is they are negatively correlated, and y (x) is leading.

Figure 5.6 shows that the wavelet coherence between ELP and RES is persistently significant at the 2-30 days period band. It is also significant and high over extended time intervals at the 30-91 days period band. The figure also shows significant coherence at the yearly period band until mid 2020, after which this seems to switch towards the 2-4 years period band. This is probably related to the structural break of the electricity price series during 2021-2023, and it could also be related to changing electricity demand patterns due to COVID policies that started in 2020. Turning to the panels with the phase differences shown on the right, one sees first that the phase differences between ELP and RES are very close to zero at the 2-30 and 30-91 days period bands, showing that both variables move in perfect lockstep at these period bands. The phase difference at the yearly period band is fairly stable and reveals an in-phase relation where the price is leading and the quantity is lagging. Here the average phase difference over the whole sample period, after conversion from radians to days, amounts to 41 days. It is unclear why here price changes are leading the quantity changes of the residual demand, perhaps this might reflect anticipatory behavior of the market participants who are well familiar with the pattern of how total electricity demand and VRE generation tends to vary across the seasons.

For the total electricity demand (TOT), the second row of Figure 5.6 shows the respective plot for the coherence with the electricity price. One can observe a significant coherence over extended time intervals only at the yearly and the 2-4 years period band. The phase differences at these period bands are fairly close to $\pm\pi/2$ (at the yearly period band, for example, the average phase difference is 83 days, while $\pi/2$ would correspond to 91 days), which makes an interpretation of the lag-lead relationship unclear.

Turning now to the third row of Figure 5.6 with the coherence between ELP and VRE generation, the overall picture looks quite similar to what is shown in the first row for ELP and RES. For the 2-14 days period band coherence is mostly significant but generally lower in comparison to what is shown in the first row, and at the yearly period band coherence is contiguously high and significant until mid 2020, after which it again switches towards the 2-4 years period band. From the relation VRE = TOT − RES, it is clear why the phase differences between ELP and VRE show approximately a mirror image of what is the case for ELP and RES, i.e. that electricity prices and VRE generation are negatively correlated. This can be seen from the arrows in the coherence plots which point to the left, revealing a phase differences close to $\pm\pi$, and this is also visible from the phase difference diagrams where the phase difference frequently switches from $(-\pi, -\pi/2)$ to $(\pi/2, \pi)$ at the lower period bands. At the yearly period band the phase difference is mostly in $(\pi/2, \pi)$, which reveals that the series move out-of-phase and that the quantity of VRE generation is leading while the price is lagging. By converting the phase difference averaged over the years 2015 - mid 2022 (until which it is fairly stable) from radians to days, I obtain an estimate for the lag between VRE generation and price which amounts to 133 days.

Turning now to the first row in Figure 5.7 for the coherence between ELP and COP, one can see significant and high coherence at the yearly period band until 2020, and at the 2-4 years period band from mid 2018 onwards. For the 2-4 years period band the phase

difference is small but all the time in $(-\pi/2, 0)$, with an average of $-0.06 \cdot 2\pi$, which upon converting to days using the mean period length of 3 years implies that changes in coal prices lead changes in electricity prices on average by 65 days. The bottom row in Figure 5.7 shows a very similar picture for the coherence between ELP and NGP, and similarly, the lead of changes in gas prices with respect to changes in electricity prices at the 2-4 years period band has a time average of 55 days. This stable lag-lead relation between fuel and electricity prices is entirely plausible and may well be the consequence of contractual fixation of coal and gas procurement prices, due to which changes in procurement prices are passed on only with a lag.

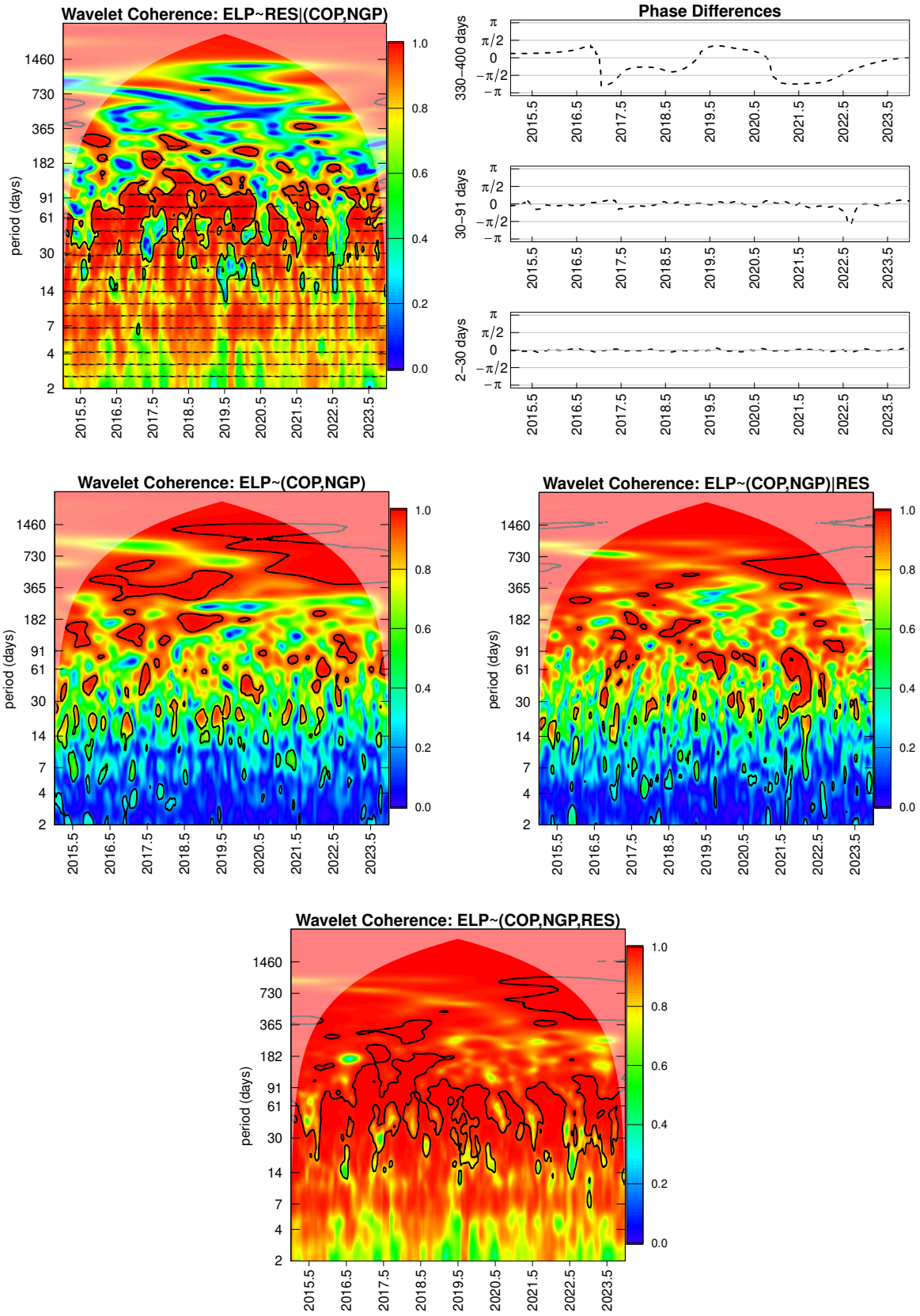
The lag length of about 60 days can be compared to a few results from related studies where a lag length can be inferred from the figures that are shown there. For the California carbon market, Aguiar-Conraria et al. (2018, p. 11) plot the phase difference between carbon and electricity prices at the yearly period band, from which I infer by inspection approximately a time average of $\pi/4$, which converts to 45 days. For the Greek electricity market Papaioannou et al. (2018, p. 52) show with a different method, namely with impulse response functions, that the response of electricity prices to a shock in gas prices reaches its full effect after about 30 days. Albeit shorter than the 60 days which I obtain for Germany, the mentioned lag lengths can be regarded as similar in magnitude, especially when considering that different countries, time periods, and variables (e.g. carbon price) are examined.

5.5.3 Multivariate Wavelet Analysis

In this section, I further disentangle the correlations between the four series (ELP, RES, COP, and NGP) by applying multivariate wavelet tools introduced in sect. 5.3.2. From the three quantity series I focus on the residual load since it neatly summarizes the influence of both the total load and VRE generation, and has turned out as the quantity series that has the highest coherence with the electricity price over a broad range of periods. Complementing the bivariate analysis from the last section with a multivariate analysis is important because there can be significant interactions between any of the series, which can also vary substantially by frequency and time. I have found in the previous section that the input prices mainly impact the electricity price at lower frequencies, while the residual load mainly impacts at high and medium frequencies. Still there is some overlap that remains, which together with interactions among the explanatory variables may have impacted the results in a significant manner. *Partial coherency analysis* allows to take care of the interactions and thus give an "unclouded" picture. Additionally, computing *multiple coherences* allows to quantify the *joint* power that multiple of the explanatory variables have for explaining the electricity price series, just like what can be assessed from the R-squared in a multivariate regression. This makes it possible to quantify the unexplained variation as a function of frequency and time.

The first row of Figure 5.8 shows the partial wavelet coherence and phase differences between ELP and RES, after controlling for the influence of both fuel price series COP and NGP. For the lower period bands (e.g., 2-91 days) both the partial coherence and the phase difference plot give a picture that is almost identical to the picture from the bivariate analysis (upper row of Figure 5.6). This reveals that, in addition to the negligible interactions of coal and gas prices with electricity prices at these lower periods, also the interactions of coal and gas prices with the residual load are negligible here. For the

Figure 5.8: Multivariate wavelet results



yearly and the 2-4 years period band, however, the partial coherence is practically never significant, which is different from what is found in the bivariate analysis. For the yearly period band the lack of a stable co-movement relationship is also visible from the phase difference diagram on the right.

The second and third row show the results for multiple coherences. Note that here no arrows are depicted since no single phase difference can be defined from multiple lag-lead relationships. The multiple wavelet coherence between electricity price and coal and gas prices is depicted on the left, while the coherence between the same variables after controlling for the influence of the residual load is depicted on the right. The third row shows the multiple wavelet coherence between electricity price, the fuel prices and the residual load. The difference between the right panel in the second row and the third row can be understood as follows. In the former, the influence of the residual load is removed from the electricity price series and from both fuel price series, so that only the variation that remains is taken into account. In the latter, the full variation of all variables is taken into account, just like for the multiple R-squared in a linear regression.

In the second row of Figure 5.8, the plots show a lot of red for all of the larger periods, which reveals how much the *joint* correlation of coal and gas prices with electricity prices is higher (and more stable) than the individual bivariate correlations. In terms of statistical significance, however, the area of significant coherence is smaller in comparison to the corresponding areas in Figure 5.7, especially in the plot on the right where the influence of the residual load is controlled for. Here significant coherence at the periods longer than 1 year exists only from mid 2020 onwards. One can conclude from this that, after removing the influence of the demand for fossil fuel powered electricity, which is the major component of the residual load, the co-movement between fossil fuel and electricity prices before mid 2020 is actually too weak to be distinguishable from "noise" (i.e., from the co-movement detectable between independent series with similar spectra as the original series). This is consistent with what can be observed from Figure 5.3 where during the price explosion all three prices move fairly close together, while for the period Jan 2015 - July 2021 changes in fuel prices co-occur only with changes in electricity prices that are of much smaller magnitude.

Lastly, the panel in the third row further illustrates how the residual load and the fuel prices have their main influence on the electricity price in distinct period bands, since the plot looks very much like a juxtaposition of the plots from the first and the second row. It also finally shows the high *joint* explanatory power that the residual load and the coal and gas prices have across the majority of the time-frequency space. In fact, the variation in electricity prices which remains unexplained exists largely at the highest frequencies, and this may simply be the inherent noise (i.e. unexplainable) component that the electricity price formation process entails.

5.6 Conclusion

In this paper, I have presented a wavelet analysis of the determinants and the periodic behavior of German wholesale electricity prices. This section summarizes the key insights and compares them with results from related studies.

First, my results confirm previous analyses that have identified the residual load as a key determinant of wholesale electricity prices (e.g., Nicolosi and Fürsch (2009); Cludius

et al. (2014); Wozabal et al. (2016); Kyritsis et al. (2017)). My results go beyond the previous analyses by determining the period band (2 days - 60 or 90 days) at which a statistically ensured in-phase co-movement relation exists between the residual load and wholesale electricity prices. I have also found that this co-movement occurs with nearly instantaneous adjustment. Conversely, there is a corresponding out-of-phase co-movement relation between VRE generation and electricity prices, although this relation is overall less persistent over time and also weaker at the highest frequencies. This is due to the fact that the residual load also captures the variation in total electricity demand, which contains an important pattern of weekly variation (see Figure 5.5 above).

Secondly, my results confirm the important role of cost fundamentals (coal and gas prices) as determinants of German wholesale electricity prices over cycles with length of a year or more. Similar results on the long-run relation between cost fundamentals and electricity prices are found in literature, usually examined with more traditional time series methods. For example, the existence of cointegration relations between electricity prices and coal, gas, or carbon prices is established by Bunn and Fezzi (2008) and Frydenberg et al. (2014) for Germany and the UK, and by Papaioannou et al. (2018) for Greece. However, in contrast to wavelet methods these traditional time series methods are limited with respect to identifying at which periodic cycles the interactions take place, and are also unable to assess how the relations between the variables change over time.[89]

Third, my results based on partial wavelet coherences reveal that the co-movement relation between cost fundamentals and electricity prices is time-varying. In post-2020 Germany electricity prices have closely followed hard coal and gas prices when these moved substantially. On the other hand, during a period of relatively stable fuel prices (2015-2020) German wholesale electricity prices have been even more stable than fuel prices, so that for 2015-2020 no statistically significant coherence can be found after controlling for the influence of the residual load. This is in line with the results of Sousa et al. (2014) and Aguiar-Conraria et al. (2018) who also find that the coherence between carbon and electricity prices is significant only for subsets of the sample period. In particular, Sousa et al. (2014, p. 125) find that the coherence stops being significant throughout the whole year 2009, and Aguiar-Conraria et al. (2018, p. 11) find that the coherence is significant only between mid 2015 and late 2016 (with a sample period covering 2014 until late 2017).

Fourth, the wavelet analysis has given insights into the lag-lead relations between fossil input prices and electricity prices. My results show that coal and gas prices lead German wholesale electricity prices by about 60 days at the lowest-frequency cycles (with cycle lengths of 2-4 years). This number is similar to the result from Aguiar-Conraria et al. (2018) for the phase difference between carbon and electricity prices of roughly 45 days at the yearly frequency, which can be inferred by visual inspection of figure 5 in Aguiar-Conraria et al. (2018, p. 11). It is also in line with the estimated impulse response function of Greek electricity prices after a gas price shock as shown by Papaioannou et al. (2018, p. 52).

Finally, the combined explanatory power of quantity variables (residual load, the difference between total demand and VRE generation) and cost fundamentals (coal and gas prices) as determinants of wholesale electricity prices has been assessed. There remains little unexplained variation across the majority of the time-frequency space (see Figure 5.8 above). When it comes to forecasting wholesale electricity prices, different market

[89] An exception is Papaioannou et al. (2018, p. 50) who employ a rolling-window Johansen test that gives insight into how the cointegration test statistic evolves over time.

participants can be interested in different time horizons and different periodicities of the electricity price movements. The analysis has shown which variables are relevant at cycles with lengths ranging from a few days to several years. For the purpose of forecasting at specific time horizons, the results have implications with regard to which variables should be included in a forecasting model for shorter and longer forecast horizons and which variables can be omitted. In particular, for forecasting at the time horizon of a week or within a season of the year (i.e. a horizon of a few months), it is essential to include total electricity demand and VRE generation, while for forecasting at horizons longer than a year, the fuel (and carbon allowance) prices paid by conventional power plants are indispensable variables. This finding can be relevant to the whole range of wholesale electricity market participants such as power generators, grid operators, large industrial electricity consumers, and regulators or government.

6 Conclusion

This chapter gives a concise summary of previous conclusions and an outlook on future research topics. The key insights from the main Chapters 3 – 5 are summarized in sect. 6.1 and the outlook follows afterwards in sect. 6.2.

6.1 Summary and Conclusions

Chapter 3 has reported bootstrap bias-corrected point and interval estimates for greenhouse gas (GHG) emission reduction potentials for 7 main sectors in a sample of 16 large European countries. The bottom line of the results is that these potentially feasible reductions from eliminating the inefficiency relative to the most efficient countries in each of seven sectors are quantitatively substantial. The estimated emission reduction potentials have been compared to the recently tightened emission reduction targets set by the European Commission (see EU (2020)). The target that is postulated for 2030 is to achieve GHG emissions amounting to only 45% of the level of emissions in 1990. With emissions in 1990 for the 16 sample countries of roughly 5000 mt, the 2030 target would require yearly emissions of 2250 mt in 2030, i.e. a reduction by 2750 mt. Since a reduction of 1000 mt or more compared to 1990 levels was already achieved by 2020, the target implies a further reduction of 1750 mt of GHG emissions until 2030. The estimates from Chapter 3 show that a substantial contribution to this from improving productive efficiency is conceivable, since the 1750 mt are not far away from the bias-corrected emission reduction potentials aggregated over both countries and sectors (ranging from 1271 mt in variant (d) to 1642 mt in variant (b)), and are well within the confidence intervals.

There are, however, several reasons why it is not realistic to expect that all of the emission reductions can be achieved in a short time period such as until 2030. These include the circumstances that structural change (e.g. tertiarization) and changes in consumer preferences are rather slow processes, that structural change is limited due to the role a country plays within the international specialization (de Araújo et al. (2020)), and that it is not uncommon that "rebound effects" occur after energy efficiency improvements have been achieved (Greening et al. (2000)).

In Chapter 4, I have formulated a stochastic model for oil production in a region, and have applied the model to discovery histories from Norway, from the shallow water part of the Gulf of Mexico (GOM Flat), and from the deep water part of the Gulf of Mexico (GOM Deep). The model contains as the key component the size-biased sampling discovery process model that is based around the idea that oil fields are not discovered as a random sample but instead via a successive sampling scheme where field size is the key variable that influences the discovery order. At first, a descriptive analysis has found that there is such a size bias for GOM Flat and Norway, while for GOM Deep the size bias is only present when analyzing all discoveries until 2020. Water depth has been found as another variable that influences the discovery order. Chapter 4 has then reported estimates of undiscovered resources and of ultimately recoverable resources (URR) derived from the parameter values that maximize the likelihood function of the discovery process model. Overall, this estimation strategy has shown the tendency to underestimate the undiscovered resources, because some of the URR estimates are below the resources known with hindsight as of 2022. The obtained resource estimates have also been compared to the published resource estimates from the official agencies of the regions (i.e. the Norwegian

Petroleum Directorate for Norway and the Bureau of Ocean Energy Management (BOEM) for GOM Deep), with the result that the majority of the obtained estimates are below the published estimates.

Then, the main purpose of Chapter 4 was to explore which curves emerge for the discovery and production rates of the overall model, and to analyze how well these curves can predict the actuals. For GOM Flat and Norway, despite implying undiscovered resources that are somewhat underestimated, the resulting curves for discovery and production match to a reasonable degree the pattern of the actuals, even quite far out-of-sample. Possible explanations for the remaining discrepancies have been outlined. For GOM Deep we have obtained less conclusive predictions, driven by large discrepancies between the estimated amounts of undiscovered resources when using discovery sequences of different lengths and in comparison to the estimates from BOEM (2017). Here the estimation based on taking the sample of discoveries until the year 2000 has returned a negative β parameter. (Recall that in the model of Chapter 4, β is the discoverability parameter where a positive value implies that oil fields tend to get smaller with the succession of discoveries, while a negative value implies an "inverse" size bias so that fields tend to get larger.) Consequentially, a fairly large undiscovered resource estimate has been obtained for this specification for GOM Deep. Although this resource estimate is close to the published point estimate from BOEM (2017), the resulting discovery rate predictions after the year 2000 are drastically above the actuals, suggesting that this high resource estimate is implausible. On the other hand, the specification with the negative β parameter has given interesting insights into conditions under which discovery and production curves that are "rather flat" can emerge, namely when this is combined with a gradual decline in the pace of new field discoveries. In contrast, our empirical results from the other regions and specifications have indicated that when the β parameter is "far enough" above zero (i.e. successive fields tend to become sufficiently smaller on average) roughly hump- or bell-shaped oil production curves emerge for the plausible scenarios of exploration well drilling and new field discoveries. In these cases oil production peaks when - or mostly before - half of the oil is extracted. This overall suggests that a 'mid-point' peak or Hubbert curve (i.e. a bell-shaped curve, see for example Campbell and Laherrère (1998); Bardi (2019); Bentley (2023)) or similar but somewhat asymmetric curves (e.g. Höök et al. (2011)) are indeed plausible approximations for the time path of oil production under frequent circumstances. Still, in such a model, a crucial parameter is of course the amount of oil that is eventually recoverable. The method for resource estimation as conducted in Chapter 4 is confronted with considerable estimation uncertainties and possible biases, also caused by the necessity to make some (hyper-)parameter choices. Nonetheless, many possible avenues for improving the procedures of oil resource estimation exist in the literature (see the last paragraphs of sect. 4.6), or possibly within the agencies that are making such estimations.

Chapter 5 has presented a wavelet analysis of the determinants and the periodic behavior of daily German wholesale electricity prices. Consistent with previous literature examining electricity markets with renewable energy, the residual load (i.e. the difference between total electricity demand and wind and solar generation) is found to be a key electricity price determinant (e.g., Nicolosi and Fürsch (2009); Cludius et al. (2014); Wozabal et al. (2016); Kyritsis et al. (2017)). By employing the powerful tools of multivariate wavelet analysis, which go beyond mere examinations of bivariate relationships, I have obtained several further insights. In particular, after removing the interaction with the coal and gas prices as confounders, it has been found that the co-movement between

the residual load and the electricity price takes place predominantly at high-frequency cycles with period lengths of 2-90 days, where the co-movement also occurs with nearly instantaneous adjustment. In this way, peaks in the corresponding cycles for solar and wind generation are associated with troughs in the cycle for electricity prices, and vice versa. Besides, the results of Chapter 5 have revealed that coal and gas prices are key determinants of German wholesale electricity prices at low-frequency cycles with period lengths above 365 days. My results have also shown that the statistical significance of this relationship is time-varying, where in particular the significance is robust only during the time period where coal and gas prices were highly volatile (2021-2023). Similar findings of time-varying relationships between energy prices exist in the literature (Sousa et al. (2014); Aguiar-Conraria et al. (2018)), which illustrates the merit of using a wavelet approach. Moreover, my results have revealed that it takes approximately 60 days for changes in coal or gas prices to be passed on as changes in electricity prices, which is also roughly in line with related estimates from the literature (Aguiar-Conraria et al. (2018); Papaioannou et al. (2018)).

Finally, Chapter 5 has shown that there is little unexplained variation in electricity prices across the time-frequency space when both quantity and price variables are used as explanatory variables, but also that there is substantial variation across the time-frequency space as to which of the explanatory variables are most important. This result can guide the construction of forecasting models of wholesale electricity prices because it provides valuable insights into how the different explanatory variables are of different importance at different periodicities. In particular, when the goal is to forecast at the time horizon of a week or within a season of the year (i.e. a horizon of a few months), it is essential to include total electricity demand and solar and wind generation, while when the goal is to forecast at horizons longer than a year, the fuel (and carbon allowance) prices are indispensable variables.

6.2 Outlook

Starting with Chapter 3, the analysis presented there has the limitation that it is purely static and does not take into account the emission reducing effects of technological change. New technologies enabling less emission-intensive forms of energy generation and mobility are crucial to reduce GHG emissions. Thus, a natural extension of the analysis presented in Chapter 3 would be the application of a dynamic method that can account for changes in the technology set occurring over time. Another extension would be to use a global country sample. Both of these possibilities have been taken up by Fait et al. (2022), where trend functions are employed to project into the future the GHG emission reduction potentials that are estimated for the period 1990-2017 and for a global country sample. A further refinement here could be the application of the Malmquist-Luenberger decomposition to separate the yearly changes in the technology set from the yearly changes in the efficiency of the countries relative to the current technology set.

In Chapter 4, the size-based sampling discovery process model that was used to estimate undiscovered resources and predict future discoveries could be refined by using other discovery process models available in the literature, with the potential for improved estimation results or simpler computations. The multivariate discovery process model as employed by Nair and Wang (1989), Lee (2008, pp. 40ff.), or Chen and Osadetz (2009) is a promising approach for delivering better estimation results. With this approach one

can go beyond the pure bivariate relation between size and discovery order and explicitly take into account other field-level characteristics that possibly influence the discovery order. Alternatively, the creaming discovery process model (see Meisner and Demirmen (1981); Lillestøl and Sinding-Larsen (2017; 2018)) is promising as it is computationally much simpler, which helps in particular in the bottom-up framework of Chapter 4 where the computations from several submodels need to combined to arrive at the final predictions for oil production, which necessarily involves a high degree of complexity. Besides, the creaming model avoids the issue with the specification of the parameter for the total number of fields prior to the parameter estimation stage (see also the discussion at the end of sect. 4.6).

Furthermore, in the model of Chapter 4 it was assumed for simplicity that the duration between the discovery of an oil field and the date of initial oil production is constant and identical for all fields. In reality, there is large variation across oil fields regarding this parameter, which can also change over time as experience from previous development projects is gained and with the adoption of new technologies. Therefore, further studies on the duration between oil field discovery and first oil production with different samples of oil fields from across the world, as well as the formulation of a predictive model for this parameter, are fruitful avenues for future research. It may also be interesting to examine in this context whether there are policies that could incentivize oil-producing firms to postpone the development specifically of some of the giant fields, or to temporally extend the initial or plateau phases of the production profiles of giant fields. Such an incentivized delay or extension of production for some of the best (i.e. largest) fields would compensate for their early discovery caused by the size bias.

However, with respect to the current global situation where many of the world's giant fields and all but a few of the world's supergiant fields are already discovered and are producing, it would first and foremost be crucial to have better data availability at a global level for these fields on the remaining recoverable reserves, production history, and other key variables. These numbers could then be used to make forecasts of the future production paths for these giants, using more sophisticated models than the simple production profiles used in Chapter 4, including for example the potential for enhanced oil recovery projects. Here also the impact of economic variables, technology options, and policy on the production profiles are important research topics that would be worth studying. When conducted at the global level, such projections for the known giant and supergiant oil fields would considerably help to assess the future prospects of oil availability and affordability.

Finally, the model of Chapter 4 could also be applied to assess the future availability of other important resources than oil such as coal, uranium, or copper. Copper in particular is a key metal because it is used in all kinds of electrical equipment and thus is vital for renewable energy expansion. For mineral ores like uranium or copper, a key variable related to deposit size is the concentration of the ore, called the ore grade. As this variable is likely to influence the order of discovery (or of first production) in addition to the variable total deposit volume (or area), it could be advantageous here to combine both variables in a multivariate discovery process model.

Chapter 5 has pointed out the differential importance of wind and solar electricity generation, total electricity demand, and fuel prices for explaining wholesale electricity prices at different time horizons (e.g. a week, several months, a year). Of course, the method as applied in Chapter 5 can equally be applied to similar data from other countries. Suitable

time series models with seasonal components could then be constructed and applied to forecast wholesale electricity prices.

A limitation of the analysis in Chapter 5 is that only wholesale electricity prices are examined (which are set at the European Energy Exchange), while the expenditures of most electricity consumers are based on the retail electricity prices as set by electricity providers. The retail prices contain as a markup the transmission grid fees and other levies like the German "EEG-Umlage".[90] An examination of the impact of rising wind and solar generation on retail electricity prices necessitates a long-run perspective, and a different method more suitable to estimate long-run relationships (since the wavelet method used in Chapter 5 is only suitable to uncover the oscillations at different frequencies). This would also require longer time series and / or panel datasets, so that data availability issues due to the short time series on wind and solar generation and also limited data availability for retail electricity prices constrain the possibility of such an analysis.

Instead, it is vital that future research on the transition away from fossil energy (including government plans and reports such as those from the EU, see e.g. EU (2019)) examines the inherent issues of a strongly increasing renewable energy share with greater realism and from a systemic perspective, as these issues suggest that the cost are nonlinear functions of the share of wind and solar generation, rising when these are employed to be the main energy sources.

As a first example, consider how the cost of integrating intermittent renewables into the electricity grid while assuring grid stability rises with their share in electricity generation due to the intermittency. That the intermittent generation can pose severe problems is for example illustrated in an analysis by Mearns and Sornette (2023) of the recent Swiss energy transition plans. The Swiss Energy Strategy 2050 was endorsed by the majority of voters in a referendum on 27$^{\text{th}}$ May of 2017.[91] The strategy entails the target to phase out all of the existing nuclear and fossil power plants and to substitute the lost power production with renewables, mainly with solar photovoltaic. The bottom line of the analysis by Mearns and Sornette (2023) is the finding that while the plan would indeed work for a summer month like July, there would be a large deficit of electricity generation for a winter month like January so that the majority of Swiss electricity demand would have to be covered by imports.

To give a second example, a battery storage is often proposed as the solution for the intermittency problem (for example in the Swiss plans, see Mearns and Sornette (2023, p. 2)). The feasibility of such a grid-scale battery storage using lithium-ion technology has been examined by Michaux (2021; 2023) in terms of the required volumes of metals. The author finds that for each of the metals copper, nickel, lithium, cobalt, and graphite the estimated required volumes vastly exceed the current mining production and the stated current reserves. Michaux (2023) concludes by stressing the need to further develop alternative battery chemistries, and also stresses the need to reassess some the existing energy transition plans as these are confronted with serious challenges.

Overall, the conclusions and examples from above highlight that a systemic perspective is vital. It is not only important to consider GHG emissions, energy, economic output

[90] Although the "EEG-Umlage" is since July 2022 paid for by the German government, this is ultimately paid for by the tax payer, so the cost of this subsidy needs to be accounted for when examining the full impact of renewable generation share on electricity generation costs.

[91] See https://www.swissinfo.ch/eng/politics/vote-opener/43190762

and the relations between these variables (see Chapter 2), but in detail all the technological options for energy supply with their practicalities, logistical bottlenecks, and environmental impacts, from a perspective not only restricted to the monetary cost but also accounting for the material and energy requirements to manufacture the technology units, as well as for the limits to growth and to the resource throughput imposed on the economic subsystem by the finite planet earth as the source of resources and the sink for waste streams.

References

Abate, J. and W. Whitt (1995): "Numerical inversion of Laplace transforms of probability distributions," *ORSA Journal on Computing*, 7 (1), 36–43.

Adelman, M. A. (1997): "My education in mineral (especially oil) economics," *Annual Review of Energy and the Environment*, 22, 13–46.

Adelman, M. A. and M. C. Lynch (1997): "Fixed view of resource limits creates undue pessimism," *Oil and Gas Journal*, 95 (14).

Aghion, P., A. Dechezleprêtre, D. Hemous, R. Martin, and J. Van Reenen (2016): "Carbon taxes, path dependency, and directed technical change: evidence from the auto industry," *Journal of Political Economy*, 124, 1–51.

Aguiar-Conraria, L., G. Conceição, and M. J. Soares (2021): "How far is gas from becoming a global commodity?" NIPE Working Papers 06/2021, NIPE - Universidade do Minho.

Aguiar-Conraria, L., M. Joana Soares, and R. Sousa (2018): "California's carbon market and energy prices: a wavelet analysis," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376, 20170256.

Aguiar-Conraria, L., M. M. Martins, and M. J. Soares (2012): "The yield curve and the macro-economy across time and frequencies," *Journal of Economic Dynamics and Control*, 36, 1950–1970.

——— (2020): "Okun's Law across time and frequencies," *Journal of Economic Dynamics and Control*, 116, 103897.

Aguiar-Conraria, L. and M. J. Soares (2014): "The continuous wavelet transform: moving beyond uni-and bivariate analysis," *Journal of Economic Surveys*, 28, 344–375.

Aldy, J. E., A. J. Krupnick, R. G. Newell, I. W. Parry, and W. A. Pizer (2010): "Designing climate mitigation policy," *Journal of Economic Literature*, 48, 903–934.

Alvarez-Ramirez, J., E. Rodriguez, and G. Espinosa-Paredes (2012): "A partisan effect in the efficiency of the US stock market," *Physica A: Statistical Mechanics and its Applications*, 391, 4923–4932.

Anderson, S. T., R. Kellogg, and S. W. Salant (2018): "Hotelling under pressure," *Journal of Political Economy*, 126 (3), 984–1026.

Arachchige, C. N., L. A. Prendergast, and R. G. Staudte (2022): "Robust analogs to the coefficient of variation," *Journal of Applied Statistics*, 49, 268–290.

Arezki, R., V. A. Ramey, and L. Sheng (2017): "News shocks in open economies: evidence from giant oil discoveries," *The Quarterly Journal of Economics*, 132 (1), 103–155.

Arps, J. J. and T. G. Roberts (1958): "Economics of drilling for Cretaceous oil on east flank of Denver-Julesburg basin," *AAPG Bulletin*, 42 (11), 2549–2566.

Asche, F., P. Osmundsen, and M. Sandsmark (2006): "The UK market for natural gas, oil and electricity: are the prices decoupled?" *The Energy Journal*, 27, 27–40.

Ayres, R. U. and B. Warr (2009): *The Economic Growth Engine: How Energy and Work Drive Material Prosperity*, Edward Elgar.

Baddeley, A. (2018): *A statistical commentary on mineral prospectivity analysis*, Springer International Publishing, chap. Handbook of Mathematical Geosciences: Fifty Years of IAMG, 25–65.

Ball, L. M., D. Leigh, and P. Mishra (2022): "Understanding U.S. inflation during the covid era," *IMF Working Paper 22/208*.

Banker, R. D., A. Charnes, and W. W. Cooper (1984): "Some models for estimating technical and scale inefficiencies in data envelopment analysis," *Management Science*, 30, 1078–1092.

Bardi, U. (2005): "The mineral economy: a model for the shape of oil production curves," *Energy Policy*, 33 (1), 53–61.

——— (2019): "Peak oil, 20 years later: failed prediction or useful insight," *Energy Research & Social Science*, 48, 257–261.

Bardi, U. and A. Lavacchi (2009): "A simple interpretation of Hubbert's model of resource exploitation," *Energies*, 2 (3), 646–661.

Bardi, U., A. Lavacchi, and L. Yaxley (2011): "Modelling EROEI and net energy in the exploitation of non renewable resources," *Ecological Modelling*, 223 (1), 54–58.

Barouch, E. and G. Kaufman (1976a): "Oil and gas discovery modelled as sampling proportional to random size," *Working Paper Alfred P. Sloan School of Management*.

——— (1976b): "Stochastic modelling of natural resource discovery - the case of oil and gas," *A talk for Colloque International Sur L'Analyse de Systemes et ses Orientations Nouvelles. December 22, 1976*.

Barouch, E., G. Kaufman, and J. Nelligan (1983): "Estimation of parameters of oil and gas discovery process models using the expectation-maximization algorithm," *Energy Modelling and Simulation*, 109–117.

Bartlett, A. A. (2000): "An analysis of US and world oil production patterns using Hubbert-style curves," *Mathematical Geology*, 32 (1), 1–17.

Benhmad, F. and J. Percebois (2018): "Photovoltaic and wind power feed-in impact on electricity prices: the case of Germany," *Energy Policy*, 119, 317–326.

Bentley, R. W. (2002): "Global oil & gas depletion: an overview," *Energy Policy*, 30 (3), 189–205.

——— (2023): "Colin Campbell, oil exploration geologist and key proponent of 'peak oil'," *Biophysical Economics and Sustainability*, 8, 3.

Best, R., P. J. Burke, and F. Jotzo (2020): "Carbon pricing efficacy: cross-country evidence," *Environmental and Resource Economics*, 77, 69–94.

Bhattacharya, P. (1974): "Convergence of sample paths of normalized sums of induced order statistics," *The Annals of Statistics*, 2 (5), 1034–1039.

Blanchard, O. J. and B. S. Bernanke (2023): "What caused the U.S. pandemic-era inflation?" NBER Working Paper 31417, National Bureau of Economic Research.

BOEM (2017): "2016a National assessment of undiscovered oil and gas resources of the U.S. outer continental shelf," Tech. rep., U.S. Department of the Interior, Bureau of Ocean Energy Management. OCS Report BOEM 2017-085.

BP (2022): "Statistical review of world energy, 71th edition," Tech. rep.

Brandt, A. R. (2007): "Testing Hubbert," *Energy Policy*, 35 (5), 3074–3088.

——— (2010): "Review of mathematical models of future oil supply: historical overview and synthesizing critique," *Energy*, 35 (9), 3958–3974.

Bunn, D. W. and C. Fezzi (2008): *Markets for Carbon and Power Pricing in Europe*, Edward Elgar Publishing Limited, chap. 6: A vector error correction model of the interactions among gas, electricity and carbon prices: an application to the cases of Germany and the United Kingdom, 145–159.

Burgess, G. L., K. K. Cross, and E. G. Kazanis (2021): "Estimated oil and gas reserves Gulf of Mexico OCS region, December 31, 2019," Tech. rep., U.S. Department of the Interior, Bureau of Ocean Energy Management, OCS Report BOEM 2021-052.

Byman, K. (2016): "Electricity Production in Sweden - IVA's Electricity Crossroads Project," *Royal Swedish Academy of Engineering Sciences, Report*.

Camarero, M., J. Castillo-Giménez, A. J. Picazo-Tadeo, and C. Tamarit (2014): "Is eco-efficiency in greenhouse gas emissions converging among European Union countries?" *Empirical Economics*, 47, 143–168.

Campbell, C. J. and J. H. Laherrère (1998): "The end of cheap oil," *Scientific American*, 278 (3), 78–83.

Cevik, S. and K. Ninomiya (2022): "Chasing the sun and catching the wind: energy transition and electricity prices in Europe," *IMF Working Paper 22/220*.

Chambers, R. G., Y. Chung, and R. Färe (1996): "Benefit and distance functions," *Journal of Economic Theory*, 70, 407–419.

Charnes, A., W. W. Cooper, and E. Rhodes (1978): "Measuring the efficiency of decision making units," *European Journal of Operational Research*, 2, 429–444.

Chen, C.-M. and M. A. Delmas (2012): "Measuring eco-inefficiency: a new frontier approach," *Operations Research*, 60, 1064–1079.

Chen, Z. and K. G. Osadetz (2009): "Improving reservoir volumetric estimations in petroleum resource assessment using discovery process models," *Petroleum Science*, 6, 105–118.

Chen, Z. and R. Sinding-Larsen (1999): "Estimating petroleum resources using geo-anchored method - a sensitivity study," *Natural Resources Research*, 8, 49–58.

Chung, Y. H., R. Färe, and S. Grosskopf (1997): "Productivity and undesirable outputs: a directional distance function approach," *Journal of Environmental Management*, 51, 229–240.

Cludius, J., H. Hermann, F. C. Matthes, and V. Graichen (2014): "The merit order effect of wind and photovoltaic electricity generation in Germany 2008-2016: estimation and distributional implications," *Energy Economics*, 44, 302–313.

Crump, K. S. (1976): "Numerical inversion of Laplace transforms using a Fourier-series approximation," *Journal of the ACM (JACM)*, 23 (1), 89–96.

Dakpo, K. H., P. Jeanneaux, and L. Latruffe (2016): "Modelling pollution-generating technologies in performance benchmarking: recent developments, limits and future prospects in the nonparametric framework," *European Journal of Operational Research*, 250, 347–359.

Daly, H. E. (2007): *Ecological Economics and Sustainable Development, Selected Esays of Herman Daly*, Edward Elgar.

Daubechies, I. (1992): *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics (vol. 61), Philadelphia: SIAM.

David, H. and J. Galambos (1974): "The asymptotic theory of concomitants of order statistics," *Journal of Applied Probability*, 11 (4), 762–770.

Day, J. W., C. F. D'Elia, A. R. Wiegman, J. S. Rutherford, C. A. Hall, R. R. Lane, and D. E. Dismukes (2018): "The energy pillars of society: perverse interactions of human resource use, the economy, and environmental degradation," *Biophysical Economics and Resource Quality*, 3, 1–16.

de Araújo, I. F., R. W. Jackson, A. B. F. Neto, and F. S. Perobelli (2020): "European Union membership and CO_2 emissions: a structural decomposition analysis," *Structural Change and Economic Dynamics*, 55, 190–203.

De Jong, C. and S. Schneider (2009): "Cointegration between gas and power spot prices," *The Journal of Energy Markets*, 2, 27–46.

de Lagarde, C. M. and F. Lantz (2018): "How renewable production depresses electricity prices: evidence from the German market," *Energy Policy*, 117, 263–277.

Debelke, J. and P. Vis (2015): *EU Climate Policy Explained*, London: Rouledge.

Delreux, T. and F. Ohler (2019): "Climate policy in European Union politics," in *Oxford Research Encyclopedia of Politics*, ed. by L. F., Oxford: Oxford University Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1–38.

di Giovanni, J., S. Kalemli-Özcan, A. Silva, and M. A. Yildirim (2023): "Pandemic-Era inflation drivers and global spillovers," Working Paper 31887, National Bureau of Economic Research.

Dickey, D. A. and W. A. Fuller (1979): "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, 74, 427–431.

Do, L. P. C., Š. Lyócsa, and P. Molnár (2019): "Impact of wind and solar production on electricity prices: quantile regression approach," *Journal of the Operational Research Society*, 70, 1752–1768.

Domazlicky, B. R. and W. L. Weber (2004): "Does environmental protection lead to slower productivity growth in the chemical industry?" *Environmental and Resource Economics*, 28, 301–324.

Dubner, H. and J. Abate (1968): "Numerical inversion of Laplace transforms by relating them to the finite Fourier cosine transform," *Journal of the ACM (JACM)*, 15 (1), 115–123.

Durbin, F. (1974): "Numerical inversion of Laplace transforms: an efficient improvement to Dubner and Abate's method," *The Computer Journal*, 17 (4), 371–376.

Ebrahimi, M. and N. C. Ghasabani (2015): "Forecasting OPEC crude oil production using a variant multicyclic Hubbert model," *Journal of Petroleum Science and Engineering*, 133, 818–823.

EEA (2019): "Trends and Projections in Europe 2019," *European Environmental Agency, EEA Report, No 15/2019*.

Elliott, G., T. J. Rothenberg, and J. H. Stock (1996): "Efficient tests for an autoregressive unit root," *Econometrica*, 64, 813–836.

Emery, G. W. and Q. Liu (2002): "An analysis of the relationship between electricity and natural-gas futures prices," *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 22, 95–122.

Energy Institute (2023): "Statistical review of world energy, 72nd edition," Tech. rep.

EU (2011): "A roadmap for moving to a low-carbon economy in 2050," *European Commission, COM(2011) 112 final, Brussels*.

——— (2018): "Going climate-neutral by 2050 - A European strategic long-term vision for a prosperous, modern, competitive and climate neutral economy," *European Commission, COM(2018) 773 final, Brussels*.

——— (2019): "Going climate-neutral by 2050: a strategic long-term vision for a prosperous, modern, competitive and climate-neutral EU economy," Tech. rep., European Commission, Directorate-General for Climate Action, Publications Office.

——— (2020): "Stepping up Europe's 2030 climate ambition," *European Commission, COM(2020) 562 final, Brussels*.

Eurostat (2015): "Manual for Air Emission Accounts, 2015 Edition," *Luxembourg: Publications Office of the European Union*.

Fait, L., J. J. Krüger, M. Tarach, and H. Wetzel (2022): "Trend projections of greenhouse gas emission reduction potentials: a bootstrap-based nonparametric efficiency analysis," *SSRN Working Paper 4107819*.

Färe, R. and S. Grosskopf (2004): *New Directions: Efficiency and Productivity*, vol. 3, New York: Springer.

Färe, R., S. Grosskopf, and F. Hernandez-Sancho (2004): "Environmental performance: an index number approach," *Resource and Energy Economics*, 26, 343–352.

Färe, R., S. Grosskopf, D.-W. Noh, and W. Weber (2005): "Characteristics of a polluting technology: theory and practice," *Journal of Econometrics*, 126, 469–492.

Färe, R., S. Grosskopf, and C. A. Pasurka Jr (2007): "Environmental production functions and environmental directional distance functions," *Energy*, 32, 1055–1066.

Färe, R., S. Grosskopf, and G. Whittaker (2013): "Directional output distance functions: endogenous directions based on exogenous normalization constraints," *Journal of Productivity Analysis*, 40, 267–269.

Färe, R. and D. Primont (1995): *Multi-Output Production and Duality: Theory and Applications*, New York: Springer.

Feenstra, R. C., R. Inklaar, and M. P. Timmer (2015): "The next generation of the Penn World Table," *American Economic Review*, 105, 3150–82.

Feigin, P. D. (1979): "On the characterization of point processes with the order statistic property," *Journal of Applied Probability*, 16 (2), 297–304.

Flor, M. A. and T. Klarl (2017): "On the cyclicity of regional house prices: new evidence for U.S. metropolitan statistical areas," *Journal of Economic Dynamics and Control*, 77, 134–156.

Fourastié, J. (1949): *Le grand espoir du XXème siècle*, Paris: Presses Universitaires de France.

Frydenberg, S., J. I. Onochie, S. Westgaard, N. Midtsund, and H. Ueland (2014): "Long-term relationships between electricity and oil, gas and coal future prices - evidence from Nordic countries, Continental Europe and the United Kingdom," *OPEC Energy Review*, 38, 216–242.

Gallegati, M., M. Gallegati, J. B. Ramsey, and W. Semmler (2011): "The US wage Phillips curve across frequencies and over time," *Oxford Bulletin of Economics and Statistics*, 73, 489–508.

Gençay, R., F. Selçuk, and B. Whitcher (2005): "Multiscale systematic risk," *Journal of International Money and Finance*, 24, 55–70.

Georgescu-Roegen, N. (1971): *The Entropy Law and the Economic Process*, Harvard University Press.

Gordon, L. (1983): "Successive sampling in large finite populations," *The Annals of Statistics*, 11 (2), 702–706.

Greening, L. A., D. L. Greene, and C. Difiglio (2000): "Energy efficiency and consumption – the rebound effect – a survey," *Energy Policy*, 28, 389–401.

Grossman, G. M. and A. B. Krueger (1991): "Environmental impacts of a North American free trade agreement," Working Paper 3914, National Bureau of Economic Research.

——— (1995): "Economic growth and the environment," *The Quarterly Journal of Economics*, 110, 353–377.

Grossmann, A. and J. Morlet (1984): "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM Journal on Mathematical Analysis*, 15, 723–736.

Growiec, J. (2012): "The world technology frontier: what can we learn from the US States?" *Oxford Bulletin of Economics and Statistics*, 74, 777–807.

Gütschow, J., L. Jeffery, R. Gieseke, A. Günther, R. Gebel, D. Stevens, M. Krapp, and M. Rocha (2019): "The PRIMAP-hist national historical emissions time series, V. 2.1. GFZ Data Services," .

Gütschow, J., M. L. Jeffery, R. Gieseke, R. Gebel, D. Stevens, M. Krapp, and M. Rocha (2016): "The PRIMAP-hist national historical emissions time series," *Earth System Science Data*, 8, 571–603.

Hagens, N. J. and D. J. White (2021): *Reality Blind: Integrating the Systems Science Underpinning Our Collective Futures (Vol. 1)*, Independently published.

Hall, C. A. and J. W. Day (2009): "Revisiting the limits to growth after peak oil," *American Scientist*, 97, 230–237.

Hall, C. A., J. G. Lambert, and S. B. Balogh (2014): "EROI of different fuels and the implications for society," *Energy Policy*, 64 (C), 141–152.

Hall, C. A. S., S. Balogh, and D. J. R. Murphy (2009): "What is the minimum EROI that a sustainable society must have?" *Energies*, 2 (1), 25–47.

Hall, C. A. S., D. Lindenberger, R. Kümmel, T. Kroeger, and W. Eichhorn (2001): "The need to integrate the natural sciences with economics," *BioScience*, 51, 663–673.

Hampf, B. and J. J. Krüger (2015): "Optimal directions for directional distance functions: an exploration of potential reductions of greenhouse gases," *American Journal of Agricultural Economics*, 97, 920–938.

Hannesson, R. (2024): "Green electricity prices," *Biophysical Economics and Sustainability*, 9.

Herbert, J. H. (1982): "A review and comparison of some commonly used methods of estimating petroleum resource availability," *Energy Sources*, 6 (4), 293–320.

——— (1983): "A concise mathematical statement of the relationship between the Arps/Roberts and Barouch/Kaufman models for estimating the petroleum resource base," *Energy Sources*, 7 (1), 33–42.

Holland, S. P. (2008): "Modeling peak oil," *The Energy Journal*, 29 (2), 61–79.

Holtz-Eakin, D. and T. M. Selden (1995): "Stoking the fires? CO2 emissions and economic growth," *Journal of Public Economics*, 57, 85–101.

Höök, M. (2014): "Depletion rate analysis of fields and regions: a methodological foundation," *Fuel*, 121 (4), 95–108.

Höök, M., R. Hirsch, and K. Aleklett (2009a): "Giant oil field decline rates and their influence on world oil production," *Energy Policy*, 37 (6), 2262–2272.

Höök, M., J. Li, N. Oba, and S. Snowden (2011): "Descriptive and predictive growth curves in energy system analysis," *Natural Resources Research*, 20 (2), 103–116.

Höök, M., B. Söderbergh, K. Jakobsson, and K. Aleklett (2009b): "The evolution of giant oil field production behavior," *Natural Resources Research*, 18 (1), 39–56.

Höök, M. and X. Tang (2013): "Depletion of fossil fuels and anthropogenic climate change - a review," *Energy Policy*, 52, 797–809.

Horn, M. K. (2014): "Giant oil and gas fields of the world," *https://edx.netl.doe.gov/dataset/aapg-datapages-giant-oil-and-gas-fields-of-the-world*.

Hotelling, H. (1931): "The economics of exhaustible resources," *Journal of Political Economy*, 39 (2), 137–175.

Hsiang, S. and R. E. Kopp (2018): "An economist's guide to climate change science," *Journal of Economic Perspectives*, 32, 3–32.

Hubbert, M. K. (1956): "Nuclear energy and the fossil fuels," *Drilling and Production Practice. American Petroleum Institute*, 7–25.

——— (1962): "Energy resources: a report to the Committee on Natural Resources of the National Academy of Sciences," Tech. rep., National Academy of Sciences - National Research Council, Washington, D.C.

——— (1982): *Techniques of prediction as applied to the production of oil and gas*, National Bureau of Standards, Washington, DC, 16–141.

IEA (2008): "World energy outlook," Tech. rep.

——— (2022): "World energy outlook," Tech. rep.

IPCC (2023): "Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]," Tech. rep., IPCC, Geneva, Switzerland, 184 pp.

Jakobsson, K. (2012): "Petroleum production and exploration: approaching the end of cheap oil with bottom-up modeling," Ph.D. thesis, Acta Universitatis Upsaliensis, Uppsala Universitet.

Jakobsson, K., R. Bentley, B. Söderbergh, and K. Aleklett (2012): "The end of cheap oil: bottom-up economic and geologic modeling of aggregate oil production curves," *Energy Policy*, 41, 860–870.

Jakobsson, K., B. Söderbergh, S. Snowden, and K. Aleklett (2014): "Bottom-up modeling of oil production: a review of approaches," *Energy Policy*, 64 (C), 113–123.

Jenkins, G. M. and D. G. Watts (1968): *Spectral Analysis and its Applications*, San Francisco: Holden-Day.

Karlin, S. and H. Taylor (1975): *A First Course in Stochastic Processes, 2nd edn*, New York: Acad. Press.

Kaufman, G. M. (1963): *Statistical Decision and Related Techniques in Oil and Gas Exploration*, Prentice-Hall.

——— (1993): "Statistical issues in the assessment of undiscovered oil and gas resources," *The Energy Journal*, 14 (1), 183–215.

Kaufman, G. M., Y. Balcer, and D. Kruyt (1975): "A probabilistic model of the oil and gas discovery process," in *Proceedings of the Conference on Energy Modeling and Forecasting, June 28-29, 1974*, AAPG Special Volumes.

Kaufmann, R. K. and C. J. Cleveland (2001): "Oil production in the lower 48 states: economic, geological, and institutional determinants," *The Energy Journal*, 22 (1), 27–49.

Kendall, M. G. and A. Stuart (1979): *The Advanced Theory of Statistics (Vol. 2), 4th edn*, London: Griffin.

Ketterer, J. C. (2014): "The impact of wind power generation on the electricity price in Germany," *Energy Economics*, 44, 270–280.

Kharecha, P. A. and J. E. Hansen (2008): "Implications of 'peak oil' for atmospheric CO2 and climate," *Global Biogeochemical Cycles*, 22, GB3012, doi:10.1029/2007GB003142.

Kholod, N., M. Evans, R. C. Pilcher, V. Roshchanka, F. Ruiz, M. Coté, and R. Collings (2020): "Global methane emissions from coal mining to continue growing even with declining coal production," *Journal of Cleaner Production*, 256, 120489.

Klopcic, A. L., J. Hojnik, and S. Bojnec (2022): "Do increased cross-border flow and greater competition in the market lead to lower electricity prices for end-consumers?" *The Electricity Journal*, 35, 107146.

Knopf, B., Y.-H. H. Chen, E. De Cian, H. Förster, A. Kanudia, I. Karkatsouli, I. Keppo, T. Koljonen, K. Schumacher, and D. P. Van Vuuren (2013): "Beyond 2020: strategies and costs for transforming the European energy system," *Climate Change Economics*, 4, 1340001.

Königsberger, K. (2003): *Analysis 1, 6. Auflage*, Berlin, Heidelberg: Springer-Verlag.

Koopmans, L. H. (1974): *The Spectral Analysis of Time Series*, New York: Acad. Press.

Korhonen, P. J. and M. Luptacik (2004): "Eco-efficiency analysis of power plants: an extension of data envelopment analysis," *European Journal of Operational Research*, 154, 437–446.

Kortelainen, M. (2008): "Dynamic environmental performance analysis: a Malmquist index approach," *Ecological Economics*, 64, 701–715.

Krautkraemer, J. A. (1998): "Nonrenewable resource scarcity," *Journal of Economic Literature*, 36 (4), 2065–2107.

Krautzberger, L. and H. Wetzel (2012): "Transport and CO_2: productivity growth and carbon dioxide emissions in the European commercial transport industry," *Environmental and Resource Economics*, 53, 435–454.

Krüger, J. J. (2021): "A wavelet evaluation of some leading business cycle indicators for the German economy," *Journal of Business Cycle Research*, 17, 293–319.

Krüger, J. J. and M. Neugart (2020): "Dissecting Okun's law beyond time and frequency," *Applied Economics Letters*, 28, 1744–1749.

Krüger, J. J. and M. Tarach (2020): "Greenhouse gas emission reduction potentials in Europe: a nonparametric efficiency analysis approach using sectoral data," *SSRN Working Paper 3716203*.

——— (2022): "Greenhouse gas emission reduction potentials in Europe by sector: a bootstrap-based nonparametric efficiency analysis," *Environmental and Resource Economics*, 81, 867–898.

Kuosmanen, T. and M. Kortelainen (2005): "Measuring eco-efficiency of production with data envelopment analysis," *Journal of Industrial Ecology*, 9, 59–72.

Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root?" *Journal of Econometrics*, 54, 159–178.

Kyritsis, E., J. Andersson, and A. Serletis (2017): "Electricity prices, large-scale renewable integration, and policy implications," *Energy Policy*, 101, 550–560.

Lee, P. J. (2008): *Statistical Methods for Estimating Petroleum Resources*, Oxford University Press.

Lillestøl, J. and R. Sinding-Larsen (2017): "Creaming and the likelihood of discovering additional giant petroleum fields," *Mathematical Geosciences*, 49, 67–83.

——— (2018): "Creaming - and the depletion of resources: a Bayesian data analysis," *NHH Dept. of Business and Management Science Discussion Paper No. 2017/16, Available at SSRN: http://dx.doi.org/10.2139/ssrn.3072338*.

Lin, C.-Y. C. (2009): "Insights from a simple Hotelling model of the world oil market," *Natural Resources Research*, 18 (1), 19–28.

Liu, Y., X. S. Liang, and R. H. Weisberg (2007): "Rectification of the bias in the wavelet power spectrum," *Journal of Atmospheric and Oceanic Technology*, 24, 2093–2102.

Lovelock, J. E. and L. Margulis (1974): "Atmospheric homeostasis by and for the biosphere: the gaia hypothesis," *Tellus*, 26, 2–10.

Maciejowska, K. (2020): "Assessing the impact of renewable energy sources on the electricity price level and variability - a quantile regression approach," *Energy Economics*, 85, 104532.

Maggio, G. and G. Cacciola (2009): "A variant of the Hubbert curve for world oil production forecasts," *Energy Policy*, 37 (11), 4761–4770.

Managi, S., J. J. Opaluch, D. Jin, and T. A. Grigalunas (2005): "Technological change and petroleum exploration in the Gulf of Mexico," *Energy Policy*, 33 (5), 619–632.

Maniatis, G. I. and N. T. Milonas (2022): "The impact of wind and solar power generation on the level and volatility of wholesale electricity prices in Greece," *Energy Policy*, 170, 113243.

Martenson, C. (2023): *The Crash Course (Revised Edn): An Honest Approach to Facing the Future of Our Economy, Energy, and Environment*, Wiley.

McGill, W. J. and J. Gibbon (1965): "The general-gamma distribution and reaction times," *Journal of Mathematical Psychology*, 2 (1), 1–18.

Meadows, D. H., D. L. Meadows, J. Randers, and W. W. Behrens (1972): *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*, New York: Universe Books.

Mearns, E. and D. Sornette (2023): "Are 2050 energy transition plans viable? A detailed analysis of projected Swiss electricity supply and demand in 2050," *Energy Policy*, 175, 113347.

Meisner, J. and F. Demirmen (1981): "The creaming method: a Bayesian procedure to forecast future oil and gas discoveries in mature exploration provinces," *Journal of the Royal Statistical Society: Series A (General)*, 144 (1), 1–31.

Michaelides, E. E. (2017): "A new model for the lifetime of fossil fuel resources," *Natural Resources Research*, 26 (2), 161–175.

Michaux, S. P. (2021): "Assessment of the extra capacity required of alternative energy electrical power systems to completely replace fossil fuels," Tech. rep., Geological Survey of Finland (GTK).

——— (2023): "Material supply challenges for the green transition to phase out fossil fuels," *in: SEB's The Green Bond Report, February 2023*.

Michel, B. (2011): "Oil production: a probabilistic model of the Hubbert curve," *Applied Stochastic Models in Business and Industry*, 27 (4), 434–449.

Mihalyi, D. (2021): "The long road to first oil (March 21, 2021)," *Available at SSRN: http://dx.doi.org/10.2139/ssrn.3719751*.

Miller, R. G. and S. R. Sorrell (2014): "The future of oil supply," *Phil. Trans. R. Soc. A*, 372: 20130179.

Mohr, S. H. and G. M. Evans (2008): "Peak oil: testing Hubbert's curve via theoretical modeling," *Natural Resources Research*, 17, 1–11.

Murphy, D. J., M. Raugei, M. Carbajales-Dale, and B. Rubio Estrada (2022): "Energy return on investment of major energy carriers: review and harmonization," *Sustainability*, 14, 7098.

Murty, S., R. R. Russell, and S. B. Levkoff (2012): "On modeling pollution-generating technologies," *Journal of Environmental Economics and Management*, 64, 117–135.

Myhre, G., A. Myhre, and F. Stordal (2001): "Historical evolution of radiative forcing of climate," *Atmospheric Environment*, 35, 2361–2373.

Nair, V. N. and P. C. C. Wang (1989): "Maximum likelihood estimation under a successive sampling discovery model," *Technometrics*, 31 (4), 423–436.

Nashawi, I. S., A. Malallah, and M. Al-Bisharah (2010): "Forecasting world crude oil production using multicyclic Hubbert model," *Energy & Fuels*, 24 (3), 1788–1800.

NEA-OECD (2018): "Nuclear Technology Development and Economics, Nuclear Energy Data," *NEA No. 7416*.

Nel, W. P. and C. J. Cooper (2009): "Implications of fossil fuel constraints on economic growth and global warming," *Energy Policy*, 37 (1), 166–180.

Nicolosi, M. and M. Fürsch (2009): "The impact of an increasing share of RES-E on the conventional power market - the example of Germany," *Zeitschrift für Energiewirtschaft*, 33, 246–254.

Nordhaus, W. (2019): "Climate change: the ultimate challenge for Economics," *American Economic Review*, 109, 1991–2014.

NPD (2021): "Resource accounts for the Norwegian continental shelf as per 31 December 2021, Report no. OD-02-22," Tech. rep., Norwegian Petroleum Directorate.

Oda, J., K. Akimoto, T. Tomoda, M. Nagashima, K. Wada, and F. Sano (2012): "International comparisons of energy efficiency in power, steel, and cement industries," *Energy Policy*, 44, 118–129.

OECD (2020): "Purchasing power parities (PPP)," *doi: 10.1787/1290ee5a-en*.

Panayotou, T. (1993): "Empirical tests and policy analysis of environmental degradation at different stages of economic development," Ilo working papers, ILO Working Paper, International Labour Organization.

Papaioannou, G. P., C. Dikaiakos, G. Evangelidis, P. G. Papaioannou, and D. S. Georgiadis (2015): "Co-movement analysis of Italian and Greek electricity market wholesale prices by using a wavelet approach," *Energies*, 8, 11770–11799.

Papaioannou, G. P., C. Dikaiakos, A. Stratigakos, A. Dramountanis, and A. T. Alexandridis (2018): "Using a rolling vector error correction model to model static and dynamic causal relations between electricity spot price and related fundamental factors: the case of Greek electricity market," *International Journal of Energy Economics and Policy*, 8, 38–54.

Patil, G. P. and C. R. Rao (1978): "Weighted distributions and size-biased sampling with applications to wildlife populations and human families," *Biometrics*, 34 (2), 179–189.

Percival, D. B. and A. T. Walden (2020): *Spectral Analysis for Univariate Time Series*, Cambridge University Press.

Perissi, I., A. Lavacchi, and U. Bardi (2021): "The role of energy return on energy invested (EROEI) in complex adaptive systems," *Energies*, 14 (24), 8411.

Picazo-Tadeo, A. J., M. Beltrán-Esteve, and J. A. Gómez-Limón (2012): "Assessing eco-efficiency with directional distance functions," *European Journal of Operational Research*, 220, 798–809.

Pindyck, R. S. (1978): "The optimal exploration and production of nonrenewable resources," *Journal of Political Economy*, 86 (5), 841–861.

Pitman, J. and N. M. Tran (2015): "Size-biased permutation of a finite sequence with independent and identically distributed terms," *Bernoulli*, 21 (4), 2484–2512.

Pukite, P., D. Coyne, and D. Challou (2019): *Mathematical Geoenergy: Discovery, Depletion, and Renewal*, vol. 241, John Wiley & Sons.

Puri, P. S. (1982): "On the characterization of point processes with the order statistic property without the moment condition," *Journal of Applied Probability*, 19 (1), 39–51.

Ramsey, J. B. and C. Lampart (1998a): "The decomposition of economic relationships by time scale using wavelets: expenditure and income," *Studies in Nonlinear Dynamics & Econometrics*, 3.

——— (1998b): "Decomposition of economic relationships by timescale using wavelets," *Macroeconomic Dynamics*, 2, 49–71.

Rashidi, K. and R. Farzipoor Saen (2015): "Measuring eco-efficiency based on green indicators and potentials in energy saving and undesirable output abatement," *Energy Economics*, 50, 18–26.

Raugei, M. (2019): "Net energy analysis must not compare apples and oranges," *Nature Energy*, 4, 86–88.

——— (2023): "Addressing a counterproductive dichotomy in the energy transition debate," *Biophysical Economics and Sustainability*, 8, 4.

Rehrl, T. and R. Friedrich (2006): "Modelling long-term oil price and extraction with a Hubbert approach: the LOPEX model," *Energy Policy*, 34 (15), 2413–2428.

Reynolds (2002): *Scarcity and Growth Considering Oil and Energy: An Alternative Neo-classical View*, The Edwin Mellen Press, pp. 240.

Reynolds, D. B. (1999): "The mineral economy: how prices and costs can falsely signal decreasing scarcity," *Ecological Economics*, 31 (1), 155–166.

——— (2001): "Oil exploration game with incomplete information: an experimental study," *Energy Sources*, 23, 571–578.

Richmond, A. K. and R. K. Kaufmann (2006): "Is there a turning point in the relationship between income and energy use and/or carbon emissions?" *Ecological Economics*, 56, 176–189.

Rintamäki, T., A. S. Siddiqui, and A. Salo (2017): "Does renewable energy generation decrease the volatility of electricity prices? An analysis of Denmark and Germany," *Energy Economics*, 62, 270–282.

Robert, C. P. and G. Casella (2010): *Monte Carlo Statistical Methods, 2nd edn*, New York: Springer.

Rockström, J., W. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, B. Nykvist, C. A. De Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sörlin, P. K. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. W. Corell, V. J. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen, and J. Foley (2009): "Planetary boundaries: exploring the safe operating space for humanity," *Ecology and Society*, 14 (2): 32.

Rösch, A. and H. Schmidbauer (2018): "WaveletComp 1.1: a guided tour through the R package, https://cran.r-project.org/web/packages/WaveletComp/," .

Rua, A. and L. C. Nunes (2009): "International comovement of stock market returns: a wavelet analysis," *Journal of Empirical Finance*, 16, 632–639.

Sapio, A. (2019): "Greener, more integrated, and less volatile? A quantile regression analysis of Italian wholesale electricity prices," *Energy Policy*, 126, 452–469.

Schmoker, J. W. and T. R. Klett (2000): *World Petroleum Assessment 2000*, U.S. Geological Survey (USGS), chap. RG: Estimating Potential Reserve Growth of Known (discovered) Fields: A Component of the USGS World Petroleum Assessment 2000.

Schuenemeyer, J. H. and L. J. Drew (1983): "A procedure to estimate the parent population of the size of oil and gas fields as revealed by a study of economic truncation," *Journal of the International Association for Mathematical Geology*, 15 (1), 145–161.

Sen, S. and H. Vollebergh (2018): "The effectiveness of taxing the carbon content of energy consumption," *Journal of Environmental Economics and Management*, 92, 74–99.

Sensfuss, F., M. Ragwitz, and M. Genoese (2008): "The merit-order effect: a detailed analysis of the price effect of renewable electricity generation on spot market prices in Germany," *Energy Policy*, 36, 3086–3094.

Shahbaz, M. and A. Sinha (2019): "Environmental Kuznets curve for CO2 emissions: a literature survey," *Journal of Economic Studies*, 46, 106–168.

Shephard, R. W. (1970): *Theory of Cost and Production Functions*, Princeton: Princeton University Press.

Shortle, J. F., J. M. Thompson, D. Gross, and C. M. Harris (2018): *Fundamentals of Queueing Theory, 5th edn*, Hoboken, New Jersey: John Wiley and Sons.

Simar, L., A. Vanhems, and P. W. Wilson (2012): "Statistical inference for DEA estimators of directional distances," *European Journal of Operational Research*, 220, 853–864.

Simar, L. and P. W. Wilson (1998): "Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models," *Management Science*, 44, 49–61.

——— (2008): *The Measurement of Productive Efficiency and Productivity Growth*, Oxford: Oxford University Press, chap. Statistical inference in nonparametric frontier models: recent developments and perspectives, 421–521.

——— (2011): "Estimation and inference in nonparametric frontier models: recent developments and perspectives," *Foundations and Trends in Econometrics*, 5, 183–337.

Simon, R. M., M. T. Stroot, and G. H. Weiss (1972): "Numerical inversion of Laplace transforms with application to percentage labeled mitoses experiments," *Computers and Biomedical Research*, 5 (6), 596–607.

Skinner, C., A. Gattinger, M. Krauss, H.-M. Krause, J. Mayer, M. G. Van Der Heijden, and P. Mäder (2019): "The impact of long-term organic farming on soil-derived greenhouse gas emissions," *Scientific Reports*, 9, 1–10.

Smil, V. (1998): "Future of oil: trends and surprises," *OPEC review*, 22 (4), 253–276.

——— (2015): *Power Density: A Key to Understanding Energy Sources and Uses*, Cambridge: The MIT Press.

——— (2022): *How the World Really Works: A Scientist's Guide to Our Past, Present and Future*, Penguin Books Ltd.

Smith, J. L. and G. L. Ward (1981): "Maximum likelihood estimates of the size distribution of North Sea oil fields," *Journal of the International Association for Mathematical Geology*, 13, 399–413.

Sorrell, S. and J. Speirs (2009): "UKERC review of evidence on global oil depletion: technical report 1: data sources and issues," Tech. rep., UK Energy Research Centre, London.

——— (2010): "Hubbert's legacy: a review of curve-fitting methods to estimate ultimately recoverable resources," *Natural Resources Research*, 19 (3), 209–230.

Sousa, R., L. Aguiar-Conraria, and M. J. Soares (2014): "Carbon financial markets: a time-frequency analysis of CO2 prices," *Physica A: Statistical Mechanics and Its Applications*, 414, 118–127.

Stark, D. (2008): "Peak production in an oil depletion model with triangular field profiles," *Journal of Interdisciplinary Mathematics*, 11 (5), 695–706.

Steffen, W., K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, C. A. de Wit, C. Folke, D. Gerten, J. Heinke, G. M. Mace, L. M. Persson, V. Ramanathan, B. Reyers, and S. Sörlin (2015): "Planetary boundaries: guiding human development on a changing planet," *Science*, 347, 1259855.

Stehrer, R., A. Bykova, K. Jaeger, O. Reiter, and M. Schwarzhappel (2019): "Industry Level Growth and Productivity Data with Special Focus on Intangible Assets: Report on Methodologies and Data Construction for the EU KLEMS Release 2019," *Vienna Institute for International Economic Studies*.

Stern, D. I. (2010): "Energy quality," *Ecological Economics*, 69, 1471–1478.

Stern, D. I. and A. Kander (2012): "The role of energy in the industrial revolution and modern economic growth," *Energy Journal*, 33, 125–152.

Thommen, A., S. Werner, O. Frank, J. Philipp, O. Knittelfelder, Y. Quek, K. Fahmy, A. Shevchenko, B. M. Friedrich, F. Jülicher, et al. (2019): "Body size-dependent energy storage causes Kleiber's law scaling of the metabolic rate in planarians," *Elife*, 8, e38187.

Thompson, E., S. Sorrell, and J. Speirs (2009): "UKERC review of evidence on global oil depletion: technical report 2: definition and interpretation of reserve estimates," Tech. rep., UK Energy Research Centre, London.

Tian, H., R. Xu, J. G. Canadell, R. L. Thompson, W. Winiwarter, P. Suntharalingam, E. A. Davidson, P. Ciais, R. B. Jackson, G. Janssens-Maenhout, et al. (2020): "A comprehensive quantification of global nitrous oxide sources and sinks," *Nature*, 586, 248–256.

Timmer, M. P., E. Dietzenbacher, B. Los, R. Stehrer, and G. J. De Vries (2015): "An illustrated user guide to the world input–output database: the case of global automotive production," *Review of International Economics*, 23, 575–605.

Tol, R. S. (2018): "The economic impacts of climate change," *Review of Environmental Economics and Policy*, 12, 4–25.

Torrence, C. and G. P. Compo (1998): "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, 79, 61–78.

Uhler, R. S. (1976): "Costs and supply in petroleum exploration: the case of Alberta," *Canadian Journal of Economics*, 72–90.

van den Bergh, J. and I. Savin (2021): "Impact of carbon pricing on low-carbon innovation and deep decarbonisation: controversies and path forward," *Environmental and Resource Economics*, 80, 705–715.

Van Nuffel, L., J. Gorenstein Dedecca, T. Smit, and K. Rademaekers (2018): "Sector coupling: how can it be enhanced in the EU to foster grid stability and decarbonise?" Tech. rep., European Parliament, Committee on Industry, Research and Energy.

Vardi, Y. (1982): "Nonparametric estimation in the presence of length bias," *The Annals of Statistics*, 10 (2), 616–620.

Veleda, D., R. Montagne, and M. Araujo (2012): "Cross-wavelet bias corrected by normalizing scales," *Journal of Atmospheric and Oceanic Technology*, 29, 1401–1408.

Verona, F. (2016): "Time-frequency characterization of the U.S. financial cycle," *Economics Letters*, 144, 75–79.

Wang, J., L. Feng, L. Zhao, S. Snowden, and X. Wang (2011): "A comparison of two typical multicyclic models used to forecast the world's conventional oil production," *Energy Policy*, 39 (12), 7616–7621.

West, G. B. (2017): *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*, Penguin Press.

Wilson, P. W. (1993): "Detecting outliers in deterministic nonparametric frontier models with multiple outputs," *Journal of Business and Economic Statistics*, 11, 319–323.

Winiwarter, W., L. Höglund-Isaksson, Z. Klimont, W. Schöpp, and M. Amann (2018): "Technical opportunities to reduce global anthropogenic emissions of nitrous oxide," *Environmental Research Letters*, 13, 014011.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Mass.: MIT Press.

Wozabal, D., C. Graf, and D. Hirschmann (2016): "The effect of intermittent renewables on the electricity price variance," *OR Spectrum*, 38, 687–709.

Würzburg, K., X. Labandeira, and P. Linares (2013): "Renewable generation and electricity prices: taking stock and new evidence for Germany and Austria," *Energy Economics*, 40, S159–S171, supplement Issue: Fifth Atlantic Workshop in Energy and Environmental Economics.

Zaim, O. and F. Taskin (2000): "Environmental efficiency in carbon dioxide emissions in the OECD: a non-parametric approach," *Journal of Environmental Management*, 58, 95–107.

Zhou, P. and B. Ang (2008): "Linear programming models for measuring economy-wide energy efficiency performance," *Energy Policy*, 36, 2911–2916.

Zhou, P., B. W. Ang, and K. L. Poh (2008a): "Measuring environmental performance under different environmental DEA technologies," *Energy Economics*, 30, 1–14.

Zhou, P., B. W. Ang, and K.-L. Poh (2008b): "A survey of data envelopment analysis in energy and environmental studies," *European Journal of Operational Research*, 189, 1–18.

Zofío, J. L. and A. M. Prieto (2001): "Environmental efficiency and regulatory standards: the case of CO_2 emissions from OECD industries," *Resource and Energy Economics*, 23, 63–83.

A Appendix to Chapter 3

Table A1: GHG emissions across sectors and countries

	Period 2008-2012				Period 2012-2016			
	CO_2	CH_4	N_2O	GHG	CO_2	CH_4	N_2O	GHG
Sectors:								
A	92.193	208.540	158.558	460.256	89.712	208.016	160.372	459.659
B	52.031	33.637	0.941	86.494	42.449	31.123	0.898	73.785
C	817.971	7.452	17.099	841.163	756.424	7.236	10.555	774.389
DE	1190.735	138.968	16.030	1345.091	1053.977	117.502	15.930	1183.434
F	52.972	0.079	1.073	54.123	49.755	0.067	1.045	50.885
G	74.259	0.316	0.563	75.132	67.838	0.250	0.596	68.618
H	469.594	4.549	4.685	478.748	449.482	4.335	5.210	458.907
Countries:								
AUT	51.107	6.920	3.191	61.114	43.931	6.470	3.172	53.483
BEL	73.401	8.399	6.847	88.198	66.619	7.929	6.059	80.396
CZE	90.279	14.302	4.285	108.843	80.649	13.945	4.396	99.051
DEU	654.103	56.302	36.961	747.224	641.271	54.210	37.423	732.635
DNK	78.315	7.153	5.405	90.926	64.979	6.857	5.402	77.228
ESP	230.152	37.313	14.355	283.416	203.419	35.532	15.064	254.097
FIN	52.091	5.091	4.471	61.539	44.399	4.740	4.543	53.900
FRA	229.250	57.504	40.761	327.310	207.434	54.678	39.493	299.578
GBR	386.425	62.979	20.077	467.680	341.522	51.729	19.234	412.077
GRC	74.828	9.422	5.151	89.561	61.145	9.045	4.219	74.408
ITA	304.934	43.995	16.758	365.478	255.947	41.205	15.619	312.340
IRL	33.364	12.069	6.111	51.442	32.168	12.760	6.373	51.625
NLD	142.834	18.525	8.112	170.053	138.514	17.553	8.051	164.691
POL	273.152	43.978	19.542	336.437	260.262	43.084	19.303	321.933
SVK	30.617	4.469	2.317	37.254	27.183	4.162	1.758	33.054
SWE	44.903	5.119	4.607	54.534	40.197	4.632	4.497	49.180
Total:								
Σ	2749.754	393.539	198.949	3341.008	2509.637	368.528	194.605	3069.678

Note: GHG (CO_2, CH_4,N_2O) and their totals (column GHG) are expressed in mt of CO_2e. Minor discrepancies may arise when the sum of the individuals GHG is compared to their totals. This is due to taking the sum and the median operations in different orders for the totals.

Table A2: Potential emission reduction of total GHG for the period 2008-2012 (left columns) and period 2012-2016 (right columns), variant (a)

	Total GHG Period 2008-12			Total GHG Period 2012-16		
	Actual	Esti-mate	Conf. Interval	Actual	Esti-mate	Conf. Interval
Sectors:						
A	460.256	213.643	[144.405, 292.685]	459.659	264.259	[210.028, 338.978]
B	86.494	47.723	[31.040, 71.862]	73.785	40.674	[25.127, 59.511]
C	841.163	453.659	[326.684, 605.299]	774.389	402.751	[287.109, 538.407]
DE	1345.091	641.263	[481.872, 890.818]	1183.434	580.397	[430.116, 845.409]
F	54.123	21.969	[17.803, 29.988]	50.885	14.559	[10.164, 22.237]
G	75.132	17.891	[9.874, 29.784]	68.618	14.956	[8.742, 25.026]
H	478.748	126.126	[93.834, 174.577]	458.907	130.819	[96.634, 179.703]
Countries:						
AUT	61.114	29.662	[25.144, 35.780]	53.483	27.579	[23.597, 33.000]
BEL	88.198	40.568	[27.973, 61.432]	80.396	36.432	[25.800, 50.537]
CZE	108.843	72.512	[68.170, 78.469]	99.051	75.065	[71.876, 79.836]
DEU	747.224	254.258	[77.702, 523.120]	732.635	252.501	[84.405, 534.271]
DNK	90.926	21.414	[4.333, 50.740]	77.228	19.342	[4.014, 45.480]
ESP	283.416	116.465	[73.985, 183.253]	254.097	113.490	[82.492, 153.683]
FIN	61.539	26.210	[16.878, 43.054]	53.900	23.291	[15.014, 39.776]
FRA	327.310	115.542	[52.775, 206.256]	299.578	135.333	[94.202, 197.693]
GBR	467.680	293.249	[263.718, 333.514]	412.077	252.008	[224.468, 292.130]
GRC	89.561	23.852	[3.886, 59.044]	74.408	23.218	[4.581, 56.536]
ITA	365.478	157.301	[119.860, 203.656]	312.340	151.437	[122.248, 190.379]
IRL	51.442	22.263	[14.082, 34.416]	51.625	15.396	[2.101, 34.861]
NLD	170.053	94.760	[79.141, 114.023]	164.691	65.654	[36.604, 115.461]
POL	336.437	211.488	[181.380, 252.492]	321.933	216.391	[194.421, 246.129]
SVK	37.254	21.470	[18.714, 25.195]	33.054	18.647	[15.938, 22.180]
SWE	54.534	21.231	[15.523, 29.114]	49.180	22.442	[17.679, 29.551]
To-tal:						
Σ	3341.008	1522.274	[1236.410, 1885.573]	3069.678	1448.417	[1176.830, 1801.477]

Note: Total GHG are expressed in million tons (mt) of CO_2 equivalents and reported with three digits following the decimal point. 95% percent bootstrap confidence intervals are in square brackets.

Table A3: Potential emission reduction of CO_2 and other GHG for the period 2012-2016, variant (b)

	Total GHG		CO_2			Other GHG		
	Estimate	Conf. Interval	Actual	Estimate	Conf. Interval	Actual	Estimate	Conf. Interval
Sectors:								
A	325.387	[237.312, 418.086]	89.712	54.631	[38.818, 75.084]	369.947	270.482	[197.798, 344.269]
B	34.018	[22.052, 54.283]	42.449	19.460	[12.755, 32.516]	31.335	14.527	[8.577, 27.235]
C	531.433	[366.803, 678.055]	756.424	513.651	[352.589, 656.921]	17.966	16.116	[12.732, 17.898]
DE	594.143	[436.285, 888.159]	1053.977	540.416	[394.722, 819.663]	129.456	53.675	[41.127, 72.248]
F	14.128	[8.990, 21.900]	49.755	13.337	[8.381, 20.830]	1.131	0.762	[0.587, 1.086]
G	14.854	[7.496, 27.264]	67.838	14.625	[7.364, 26.891]	0.780	0.229	[0.131, 0.422]
H	127.955	[85.080, 184.512]	449.482	125.931	[83.823, 181.453]	9.425	2.023	[1.130, 3.633]
Countries:								
AUT	29.638	[23.148, 39.612]	43.931	23.227	[17.342, 31.954]	9.552	6.411	[5.621, 7.914]
BEL	47.080	[25.862, 71.281]	66.619	40.947	[24.050, 60.221]	13.777	6.122	[1.651, 12.214]
CZE	76.153	[71.936, 82.703]	80.649	60.313	[56.710, 66.069]	18.402	15.840	[15.079, 17.177]
DEU	265.944	[85.221, 557.340]	641.271	191.177	[27.081, 455.785]	91.364	67.273	[53.497, 88.284]
DNK	17.115	[2.825, 47.659]	64.979	12.769	[2.580, 37.879]	12.249	4.337	[0.247, 10.173]
ESP	128.989	[82.070, 211.433]	203.419	107.048	[75.074, 168.102]	50.678	21.811	[6.443, 44.074]
FIN	22.655	[14.696, 40.503]	44.399	16.973	[10.183, 32.577]	9.502	5.632	[4.434, 8.367]
FRA	166.415	[95.332, 268.695]	207.434	81.502	[46.430, 137.705]	92.144	59.382	[46.109, 88.175]
GBR	312.612	[230.036, 388.475]	341.522	261.110	[186.959, 327.932]	70.555	48.821	[40.614, 66.674]
GRC	20.325	[1.994, 51.685]	61.145	15.959	[1.086, 41.210]	13.263	4.361	[0.397, 10.663]
ITA	164.720	[119.502, 212.120]	255.947	136.883	[102.482, 171.134]	56.393	27.840	[15.104, 44.591]
IRL	15.246	[2.540, 36.284]	32.168	8.243	[1.253, 20.714]	19.457	6.982	[0.210, 16.791]
NLD	72.755	[34.253, 132.484]	138.514	61.949	[31.583, 110.987]	26.178	10.794	[2.332, 22.501]
POL	236.028	[193.817, 303.716]	260.262	206.024	[174.238, 250.767]	61.671	27.304	[16.968, 51.591]
SVK	23.882	[15.266, 31.372]	27.183	19.455	[12.423, 25.303]	5.871	4.176	[2.578, 5.694]
SWE	22.244	[16.225, 30.308]	40.197	16.608	[11.477, 23.412]	8.983	5.626	[4.669, 7.620]
Total:								
Σ	1642.282	[1321.325, 2049.171]	2509.637	1282.153	[998.849, 1646.980]	560.041	360.167	[280.535, 443.626]

Note: GHG (CO_2, other GHG) and their totals are expressed in million tons (mt) of CO_2 equivalents and reported with three digits following the decimal point. 95% percent bootstrap confidence intervals are in square brackets.

Table A4: Potential emission reduction of CO_2, CH_4 and N_2O for the period 2012-2016, variant (c)

	Total GHG		CO_2			CH_4			N_2O		
	Estimate	Conf. Interval	Actual	Estimate	Conf. Interval	Actual	Estimate	Conf. Interval	Actual	Estimate	Conf. Interval
Sectors:											
A	341.330	[257.191, 425.650]	89.712	59.234	[42.863, 76.725]	208.016	143.696	[106.738, 182.262]	160.372	134.820	[103.471, 158.586]
B	36.853	[23.132, 58.477]	42.449	20.796	[13.217, 34.021]	31.123	15.370	[8.694, 28.111]	0.898	0.599	[0.407, 0.869]
C	467.694	[326.398, 610.573]	756.424	449.334	[311.293, 588.609]	7.236	6.496	[5.303, 7.211]	10.555	9.449	[7.631, 10.503]
DE	577.091	[432.825, 830.843]	1053.977	528.289	[394.903, 770.177]	117.502	40.962	[31.199, 54.817]	15.930	7.811	[5.838, 10.624]
F	12.988	[8.426, 19.625]	49.755	12.315	[7.893, 18.757]	0.067	0.024	[0.018, 0.036]	1.045	0.645	[0.494, 0.960]
G	12.608	[6.094, 23.242]	67.838	12.324	[5.910, 22.841]	0.250	0.078	[0.046, 0.126]	0.596	0.206	[0.127, 0.339]
H	37.940	[3.892, 92.966]	449.482	37.040	[3.723, 90.690]	4.335	0.354	[0.024, 1.741]	5.210	0.544	[0.145, 1.210]
Countries:											
AUT	32.156	[22.853, 45.152]	43.931	23.951	[17.068, 34.342]	6.470	4.681	[3.803, 6.290]	3.172	2.243	[1.767, 3.137]
BEL	47.297	[25.222, 71.832]	66.619	40.992	[23.300, 61.279]	7.929	2.619	[0.141, 6.063]	6.059	3.538	[1.468, 5.807]
CZE	76.383	[70.334, 84.408]	80.649	60.562	[55.347, 67.520]	13.945	12.442	[11.744, 13.399]	4.396	3.334	[3.023, 3.941]
DEU	245.496	[76.096, 498.813]	641.271	169.158	[18.291, 398.438]	54.210	37.954	[27.648, 52.652]	37.423	29.974	[25.817, 36.728]
DNK	13.531	[2.267, 45.434]	64.979	9.314	[1.990, 36.235]	6.857	2.346	[0.105, 5.386]	5.402	1.869	[0.110, 4.350]
ESP	113.582	[59.533, 186.568]	203.419	92.713	[53.212, 147.129]	35.532	14.451	[4.788, 27.122]	15.064	6.372	[1.457, 13.158]
FIN	26.336	[14.704, 44.648]	44.399	19.105	[10.089, 34.013]	4.740	2.603	[1.469, 4.304]	4.543	3.610	[3.037, 4.394]
FRA	143.806	[57.583, 246.310]	207.434	61.895	[9.513, 125.914]	54.678	35.281	[20.628, 53.518]	39.493	30.094	[25.871, 37.332]
GBR	214.855	[148.945, 305.637]	341.522	160.572	[108.855, 232.902]	51.729	36.289	[27.030, 50.561]	19.234	13.751	[11.451, 18.696]
GRC	18.977	[1.744, 46.205]	61.145	14.885	[0.840, 37.384]	9.045	2.662	[0.156, 6.454]	4.219	1.429	[0.107, 3.272]
ITA	166.767	[113.828, 255.337]	255.947	139.122	[99.935, 213.384]	41.205	20.349	[11.267, 33.504]	15.619	7.289	[2.703, 13.552]
IRL	13.628	[1.097, 33.587]	32.168	6.870	[0.647, 18.920]	12.760	4.496	[0.127, 10.618]	6.373	2.257	[0.067, 5.358]
NLD	66.267	[31.673, 120.175]	138.514	56.109	[29.483, 100.711]	17.553	6.242	[0.817, 13.765]	8.051	3.886	[1.287, 7.243]
POL	249.855	[190.148, 310.531]	260.262	214.603	[169.505, 254.559]	43.084	12.780	[4.562, 28.561]	19.303	14.423	[11.715, 18.580]
SVK	22.744	[14.832, 31.106]	27.183	18.341	[12.173, 24.952]	4.162	2.912	[1.980, 4.037]	1.758	1.076	[0.530, 1.670]
SWE	18.192	[10.291, 32.972]	40.197	11.423	[5.485, 23.201]	4.632	2.880	[2.060, 4.461]	4.497	3.266	[2.665, 4.307]
Total:											
Σ	1486.927	[1213.634, 1839.676]	2509.637	1119.317	[872.839, 1435.472]	368.528	208.709	[166.668, 253.340]	194.605	157.971	[123.828, 189.328]

Note: GHG (CO_2, CH_4,N_2O) and their totals are expressed in million tons (mt) of CO_2 equivalents and reported with three digits following the decimal point. 95% percent bootstrap confidence intervals are in square brackets. Minor discrepancies may arise when the sum of the individuals GHG is compared to their totals. This is due to taking the sum and the median operations in different orders for the totals.

Table A5: Potential emission reduction of total GHG and output enhancement for the period 2012-2016, variant (d)

	Output			Total GHG		
	Actual	Estimate	Conf. Interval	Actual	Estimate	Conf. Interval
Sectors:						
A	176.698	44.130	[30.172, 62.709]	459.659	248.062	[174.338, 336.317]
B	80.654	35.034	[19.751, 55.689]	73.785	36.179	[22.278, 56.240]
C	1809.919	678.039	[425.829, 1014.403]	774.389	306.796	[212.143, 424.805]
DE	314.613	49.079	[28.352, 85.066]	1183.434	547.370	[397.768, 792.031]
F	597.705	35.946	[12.313, 75.856]	50.885	14.100	[10.217, 21.628]
G	1320.321	94.173	[45.442, 168.895]	68.618	11.536	[7.331, 18.844]
H	546.876	47.462	[27.166, 78.382]	458.907	107.550	[82.477, 145.105]
Countries:						
AUT	126.909	28.021	[20.330, 37.812]	53.483	14.858	[11.615, 19.785]
BEL	134.940	15.215	[6.451, 31.835]	80.396	42.007	[24.473, 69.574]
CZE	118.664	44.038	[37.385, 54.093]	99.051	49.621	[42.766, 60.568]
DEU	1124.347	207.739	[26.270, 524.734]	732.635	171.658	[62.563, 360.565]
DNK	72.893	15.874	[6.789, 27.895]	77.228	10.218	[1.709, 22.885]
ESP	442.674	131.032	[99.570, 175.213]	254.097	94.736	[58.357, 165.020]
FIN	65.276	24.055	[16.677, 35.392]	53.900	16.570	[10.382, 28.677]
FRA	625.329	92.779	[36.879, 190.854]	299.578	156.579	[84.457, 243.573]
GBR	612.464	109.440	[80.558, 146.340]	412.077	238.760	[194.299, 321.503]
GRC	62.585	8.165	[1.552, 18.656]	74.408	14.299	[3.408, 34.666]
ITA	602.858	174.096	[111.289, 249.810]	312.340	96.580	[77.844, 127.095]
IRL	61.638	11.475	[1.313, 30.100]	51.625	9.361	[1.278, 21.550]
NLD	225.522	30.538	[15.204, 57.622]	164.691	60.244	[34.418, 103.644]
POL	376.336	45.867	[24.653, 80.858]	321.933	247.319	[184.337, 315.200]
SVK	62.312	43.017	[36.939, 50.859]	33.054	5.665	[4.376, 7.861]
SWE	132.039	2.511	[0.847, 5.233]	49.180	22.626	[17.129, 30.944]
Total:						
Σ	4846.786	983.863	[676.145, 1380.091]	3069.678	1271.601	[1028.342, 1594.755]

Note: Total GHG and Output are expressed in million tons (mt) of CO_2 equivalents and in bn € (in PPP of 2010), respectively. Both are reported with three digits following the decimal point. 95% percent bootstrap confidence intervals are in square brackets.

B Formal Appendix to Chapter 4

B.1 Derivation of the Posterior Distribution from a Poisson prior

Under the Poisson prior assumption of $N \sim \text{Pois}(\nu)$, Bayes' rule implies for the posterior of N that

$$\mathbb{P}[N = n|\mathcal{F}] = \frac{\mathrm{e}^{-\nu}\frac{\nu^n}{n!}\ell(\boldsymbol{x}|n)}{\sum_{n=k}^{\infty}\mathrm{e}^{-\nu}\frac{\nu^n}{n!}\ell(\boldsymbol{x}|n)}, \tag{B.1}$$

where $\ell(\boldsymbol{x}|n)$ is given in equation (4.13). Substituting $l = n - k$, the numerator in (B.1) can be rearranged as follows

$$\begin{aligned}
\mathbb{P}[N-k = l|\mathcal{F}] &\propto \mathrm{e}^{-\nu}\frac{\nu^{(k+l)}}{(k+l)!}\frac{(k+l)!}{l!}\Big(\prod_{i=1}^{k}\frac{f_{\boldsymbol{\theta}}(x_i)}{p_{\text{tr}}(\boldsymbol{\theta})}\frac{x_i^{\beta}}{b_i}\Big)\Big(\int_0^{\infty}\phi_{\boldsymbol{\theta}}(\gamma)^l g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\Big),\\
\mathbb{P}[N-k = l|\mathcal{F}] &\propto \nu^k\Big(\prod_{i=1}^{k}\frac{f_{\boldsymbol{\theta}}(x_i)}{p_{\text{tr}}(\boldsymbol{\theta})}\frac{x_i^{\beta}}{b_i}\Big)\Big(\int_0^{\infty}\mathrm{e}^{-\nu}\frac{(\nu\phi(\gamma))^l}{l!}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\Big).
\end{aligned}$$

Hence the posterior of $N-k$, the number of undiscovered fields, becomes

$$\mathbb{P}[N-k = l|\mathcal{F}] = \frac{\int_0^{\infty}\mathrm{e}^{-\nu}\cdot\frac{(\nu\phi(\gamma))^l}{l!}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}{\sum_{l=0}^{\infty}\int_0^{\infty}\mathrm{e}^{-\nu}\cdot\frac{(\nu\phi(\gamma))^l}{l!}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}.$$

We also need the posterior pdf of Γ_k given $\boldsymbol{X} = \boldsymbol{x}$ and $N - k = l$ from the first line of equation (4.14), which is

$$g^{(0)}(\gamma|\boldsymbol{x}, l) \coloneqq \frac{\phi_{\boldsymbol{\theta}}(\gamma)^l g_{\boldsymbol{b}}(\gamma)}{\int_0^{\infty}\phi_{\boldsymbol{\theta}}(\gamma)^l g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}.$$

In Lemmas 1-2, we derive a specific refactoring of the joint distribution of $(\Gamma_k, N-k)|\mathcal{F}$, which proves the claim from sect. 4.3.2 that $N-k$ has a mixed Poisson posterior. In Lemmas 3-4, we derive some further formulas from the joint distribution. The formulas in Lemma 3 for $\mathbb{E}[(N-k)^1|\mathcal{F}]$ and $g^{(1)}(\gamma|\mathcal{F})$ have already been derived by Lee (2008) (see equations (A.72)-(A.74) on pp. 195f.), while the refactoring in Lemma 2 has not been stated in the literature, at least to our knowledge.

Lemma 1. *It holds that*

$$\sum_{l=0}^{\infty}\int_0^{\infty}\mathrm{e}^{-\nu}\cdot\frac{(\nu\phi(\gamma))^l}{l!}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma = \int_0^{\infty}\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma. \tag{B.2}$$

Proof of Lemma 1: Interchanging the order of summation and integration is allowed by Fubini's theorem since all terms are positive. The remainder follows directly from the definition of the exponential function. □

Lemma 2. *The joint distribution $g^{(0)}(\gamma|\boldsymbol{x}, l)\cdot\mathbb{P}[N-k=l|\mathcal{F}]$ can be refactored as a continuous mixture of Poisson distributions. In particular,*

$$\begin{aligned}
g^{(0)}(\gamma|\boldsymbol{x}, l)\cdot\mathbb{P}[N-k=l|\mathcal{F}] &= \mathbb{P}[N-k=l|\Gamma_k=\gamma]\cdot g^{(0)}(\gamma|\mathcal{F}),\\
\text{where } \mathbb{P}[N-k=l|\Gamma_k=\gamma] &= \frac{(\nu\phi(\gamma))^l}{l!}\mathrm{e}^{-\nu\phi(\gamma)},\\
\text{and } g^{(0)}(\gamma|\mathcal{F}) &= \frac{\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)}{\int_0^{\infty}\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}.
\end{aligned} \tag{B.3}$$

Proof of Lemma 2:

$$\begin{aligned} g^{(0)}(\gamma|\boldsymbol{x},l)\cdot\mathbb{P}[N-k=l|\mathcal{F}] &= \left(\frac{\phi(\gamma)^l g_{\boldsymbol{b}}(\gamma)}{\int_0^\infty \phi(\gamma)^l g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}\right)\cdot\left(\frac{\int_0^\infty \mathrm{e}^{-\nu}\cdot\frac{(\nu\phi(\gamma))^l}{l!}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}{\int_0^\infty \mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}\right) \\ &= \phi(\gamma)^l g_{\boldsymbol{b}}(\gamma)\cdot\frac{\mathrm{e}^{-\nu}\frac{\nu^l}{l!}}{\int_0^\infty \mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma} \\ &= \frac{(\nu\phi(\gamma))^l}{l!}\cdot\frac{\mathrm{e}^{-\nu}g_{\boldsymbol{b}}(\gamma)}{\int_0^\infty \mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma} \\ &= \left(\frac{(\nu\phi(\gamma))^l}{l!}\mathrm{e}^{-\nu\phi(\gamma)}\right)\cdot\left(\frac{\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)}{\int_0^\infty \mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}\right). \end{aligned} \tag{B.4}$$

We first apply the result of Lemma 1 in the denominator of $\mathbb{P}[N-k=l|\mathcal{F}]$, then cancel the integrals $\int_0^\infty \phi(\gamma)^l g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma$, then reorder some terms, and finally expand by $\mathrm{e}^{-\nu\phi(\gamma)}$.

□

Lemma 3. *For $m=1,2$, it holds for the posterior moments $\mathbb{E}[(N-k)^m|\mathcal{F}]$ that*

$$\begin{aligned} \mathbb{E}[(N-k)^1|\mathcal{F}] &= \frac{\int_0^\infty \nu\phi(\gamma)\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}{\int_0^\infty \mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}, \\ \mathbb{E}[(N-k)^2|\mathcal{F}] &= \frac{\int_0^\infty \nu\phi(\gamma)\big(1+\nu\phi(\gamma)\big)\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}{\int_0^\infty \mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}. \end{aligned} \tag{B.5}$$

Also, for $m=1,2$, it holds for the densities

$$g^{(m)}(\gamma|\mathcal{F}) \coloneqq \Big(\sum_{l=0}^\infty l^m\mathbb{P}[N-k=l|\Gamma_k=\gamma]g^{(0)}(\gamma|\mathcal{F})\Big)\Big/\mathbb{E}[(N-k)^m|\mathcal{F}]$$

that

$$\begin{aligned} g^{(1)}(\gamma|\mathcal{F}) &= \frac{\nu\phi(\gamma)\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)}{\int_0^\infty \nu\phi(\gamma)\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}, \\ g^{(2)}(\gamma|\mathcal{F}) &= \frac{\nu\phi(\gamma)\big(1+\nu\phi(\gamma)\big)\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)}{\int_0^\infty \nu\phi(\gamma)\big(1+\nu\phi(\gamma)\big)\mathrm{e}^{\nu(\phi(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma}. \end{aligned} \tag{B.6}$$

Proof of Lemma 3: (B.5) follows from the properties of the Pois(λ)-distribution which has mean λ and second moment $\lambda(1+\lambda)$, together with the law of iterated expectations / the law of total probability. To see (B.6), note that integrating $\sum_{l=0}^\infty l^m\mathbb{P}[N-k=l|\Gamma_k=\gamma]g^{(0)}(\gamma|\mathcal{F})$ over γ gives by definition the m^{th} posterior moment from above. Thus, the numerators in (B.6) have to match the numerators in (B.5), with the only difference that the numerators in (B.6) have no integral. □

Lemma 4. *Consider any continuous transformation of Γ_k, denoted by $t^*(\Gamma_k)$. For $m=1,2$, provided that $\int_0^\infty t^*(\gamma)g^{(m)}(\gamma|\mathcal{F})\mathrm{d}\gamma$ exists, it holds that*

$$\mathbb{E}[(N-k)^m t^*(\Gamma_k)|\mathcal{F}] = \mathbb{E}[(N-k)^m|\mathcal{F}]\int_0^\infty t^*(\gamma)g^{(m)}(\gamma|\mathcal{F})\mathrm{d}\gamma. \tag{B.7}$$

Proof of Lemma 4: We write out the expectation after factoring the joint distribution as $\mathbb{P}[N-k=l|\Gamma_k=\gamma]\cdot g^{(0)}(\gamma|\mathcal{F})$, then pull $t^*(\gamma)$ out of the sum, then expand by $\mathbb{E}[(N-k)^m|\mathcal{F}]$, and finally recognize the formula of $g^{(m)}(\gamma|\mathcal{F})$. □

B.2 Numerical Computation of the General Gamma Density

A draw from the general gamma distribution $g_{\boldsymbol{b}}(\gamma)$ with parameters $\boldsymbol{b} = (b_1, ..., b_k)$ and support $[0, \infty)$ as defined in sect. 4.3.2 is easily obtained by the sum $\sum_{j=1}^{k} \varepsilon_j / b_j$, where ε_j are iid standard exponential variates. With this one can approximate the density $g_{\boldsymbol{b}}(\gamma)$ by Monte Carlo methods, and one could try to approximate the integrals of the type $\int_0^\infty \mathrm{e}^{\nu(\phi(\gamma)-1)} g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma$ or $\int_0^\infty \phi(\gamma)^{n-k} g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma$ by Monte Carlo integration. However, the integrands $\mathrm{e}^{-\nu[1-\phi(\gamma)]}$ or $\phi(\gamma)^{n-k}$ exhibit a very sharp decline after zero due to the large exponents ν and $n-k$.[92] Thus, a highly precise evaluation of $g_{\boldsymbol{b}}(\gamma)$ is required at its left tail in the proximity of zero. Since Monte Carlo methods do not perform well at approximating the tails of a distribution, they have not been used in the literature here to our knowledge.

The pdf $g_{\boldsymbol{b}}(\gamma)$ does have an analytical expression, which can be obtained via partial fractions expansion from its real-valued Laplace transform that is given by

$$\psi_{\boldsymbol{b}}(s) \coloneqq \prod_{j=1}^{k} \frac{b_j}{b_j + s}, \; s \in \mathbb{R}_+. \tag{B.8}$$

The resulting expression is (McGill and Gibbon (1965), pp. 4f.; Barouch and Kaufman (1976a), p. 13):

$$g_{\boldsymbol{b}}(\gamma) = \sum_{j=1}^{k} C_j b_j \mathrm{e}^{-\gamma b_j}, \quad \text{where } C_j = \prod_{l=1, l\neq j}^{k} \frac{b_l}{b_l - b_j} \text{ for } j = 1, \ldots, k.$$

However, this expression is practically useless for accurately evaluating the density, in particular at the important points near zero (Nair and Wang (1989), pp. 430f.). The problem with this formula is that already for modest values of k the coefficients C_j, $j = 1, ..., k$, become extremely large and are always alternating in sign so that each coefficient would need to be computed with numerically infeasible precision.

In the literature, the method that is instead used to compute $g_{\boldsymbol{b}}(\gamma)$ is the Fourier-series method for numerically inverting the complex-valued Laplace transform of $g_{\boldsymbol{b}}(\gamma)$. Let $\mathrm{i} = \sqrt{-1}$ denote the imaginary number, and insert the complex number $s = a + \mathrm{i}\omega$ into (B.8). The inversion integral $\psi_{\boldsymbol{b}}(s) \to g_{\boldsymbol{b}}(\gamma)$ of the Laplace transform $g_{\boldsymbol{b}}(\gamma) \to \psi_{\boldsymbol{b}}(s)$ is given by either of the following formulas (Abate and Whitt (1995), p. 37)

$$\begin{aligned} g_{\boldsymbol{b}}(\gamma) &= \frac{\mathrm{e}^{a\gamma}}{\pi} \int_0^\infty \big[\mathrm{Re}\{\psi_{\boldsymbol{b}}(a+\mathrm{i}\omega)\}\cos(\omega\gamma) - \mathrm{Im}\{\psi_{\boldsymbol{b}}(a+\mathrm{i}\omega)\}\sin(\omega\gamma)\big]\mathrm{d}\omega, \; \gamma \in \mathbb{R}, \\ g_{\boldsymbol{b}}(\gamma) &= \frac{2\mathrm{e}^{a\gamma}}{\pi} \int_0^\infty \mathrm{Re}\{\psi_{\boldsymbol{b}}(a+\mathrm{i}\omega)\}\cos(\omega\gamma)\mathrm{d}\omega, \; \gamma \in \mathbb{R}_+, \end{aligned} \tag{B.9}$$

where a can be an arbitrary real number greater than the real parts of all singularities of the complex-valued Laplace transform. For the general gamma density this requirement implies that $a > -b_k$.

[92] The Laplace transform $\phi(\gamma)$ fulfills $\phi(0) = 1$, is convex, and declines monotonously towards the asymptote $\lim_{\gamma\to\infty} \phi(\gamma) = 0$. This carries over to the integrand $\phi(\gamma)^{n-k}$. The integrand $\mathrm{e}^{-\nu[1-\phi(\gamma)]}$ also starts at 1, is convex, and declines monotonously towards the asymptote $\mathrm{e}^{-\nu}$. Both integrands exhibit an extremely sharp decline in the proximity of zero.

Inverting Laplace transforms via the Fourier-series method means essentially that a trapezoidal quadrature rule is applied to either of the integrals from (B.9). The key result is that a carefully chosen trapezoidal rule is equivalent to a Fourier-series approximation of the density of interest on a certain compact interval (Dubner and Abate (1968); Durbin (1974); Crump (1976); Abate and Whitt (1995)). This makes the method particular effective as a numerical technique.

The version of the Fourier-series method we found most useful and practical to implement for evaluating the general gamma density is the "Euler method". For probability applications this method is neatly outlined by Abate and Whitt (1995). In particular, here the trapezoidal rule is applied to the second line of (B.9) with a step-size of $h = \pi/(2\gamma)$. Denoting the Fourier-series approximation of $g_{\boldsymbol{b}}(\gamma)$ by $\widetilde{g}_{\boldsymbol{b}}(\gamma)$, and also setting $a = A/(2\gamma)$, the approximation at each γ is the nearly alternating series (Abate and Whitt (1995), pp. 37f.)

$$\widetilde{g}_{\boldsymbol{b}}(\gamma) = \frac{\mathrm{e}^{A/2}}{\gamma}\left(\frac{1}{2}\psi_{\boldsymbol{b}}\Big(\frac{A}{2\gamma}\Big) + \sum_{m=1}^{\infty}(-1)^m \mathrm{Re}\Big\{\psi_{\boldsymbol{b}}\Big(\frac{A + \mathrm{i}(2m\pi)}{2\gamma}\Big)\Big\}\right). \tag{B.10}$$

For the general gamma distribution, the real part of the complex-valued Laplace transform is easily derived from (B.8) as

$$\mathrm{Re}\{\psi_{\boldsymbol{b}}(a + \mathrm{i}\omega)\} = \Big(\prod_{j=1}^{k} \frac{b_j}{\sqrt{(b_j + a)^2 + \omega^2}}\Big) \cdot \cos\Big(\sum_{j=1}^{k} \mathrm{atan}\big(\frac{\omega}{b_j + a}\big)\Big). \tag{B.11}$$

It can be shown generally that the discretization error of the approximation $\widetilde{g}_{\boldsymbol{b}}(\gamma)$, i.e. the difference between the series in (B.10) and the exact integral in (B.9), equals (Abate and Whitt (1995), p. 38)

$$\widetilde{g}_{\boldsymbol{b}}(\gamma) - g_{\boldsymbol{b}}(\gamma) = \sum_{m=1}^{\infty} \mathrm{e}^{-mA} g_{\boldsymbol{b}}(\gamma(2m + 1)). \tag{B.12}$$

The crux of the method is that the free parameter A can always be chosen so that the discretization error becomes as small as desired. For example, if we wanted a discretization error of exactly 10^{-E} at some γ, we would need to pick A so that the right-hand side of (B.12) equals 10^{-E}. Obviously, since $g_{\boldsymbol{b}}(\gamma(2m + 1))$ is not known this cannot be done. What can be done is bounding (B.12) by a function that envelopes $g_{\boldsymbol{b}}(\gamma)$, and then computing A based on this envelope so that the discretization error is forced to be less than 10^{-E}.

The delicacy of the Fourier-series method is that there is a risk of setting A too high, which can lead to large error. The reason is that the discretization error (B.12) is not the only error, there is also an error from the necessity to truncate the series in (B.10). The term $\mathrm{e}^{A/2}$ reveals that this truncation error grows exponentially with A. To mitigate the problem, Euler summation for accelerating the convergence of the series has been proposed by Simon et al. (1972), which allows to approximate the limit of a nearly alternating series such as (B.10) much faster. According to Abate and Whitt (1995), Euler summation for a nearly alternating series is equivalent to a simple weighted average of the last ΔM partial sums, where the weights are from the binomial distribution with parameters ΔM and $p = 1/2$. Put more clearly, Euler summation of (B.10) amounts to computing the terms inside the sum of (B.10) for $m = M, M + 1, \ldots, M + \Delta M$, then computing the

corresponding partial sums while also adding the initial M terms, and finally averaging with the binomial probabilities.[93]

The remainder of this section is concerned with how to envelope the function $g_{\boldsymbol{b}}(\gamma)$ in order to determine A as to achieve the desired discretization error 10^{-E}.[94] The usual candidate for bounding $g_{\boldsymbol{b}}(\gamma)$ is its maximum. As explained by McGill and Gibbon (1965), it is clear that $g_{\boldsymbol{b}}(\gamma)$ behaves globally similar to a usual gamma density, in particular that $g_{\boldsymbol{b}}(0) = 0$ when $k \geq 2$, $\lim_{\gamma\to\infty} g_{\boldsymbol{b}}(\gamma) = 0$, and that the function has a unique maximum. The maximum is not derivable analytically, but is easily approximated by a Monte Carlo simulation.[95]

In the following, we propose a novel idea for how to get an improved envelope for $g_{\boldsymbol{b}}(\gamma)$ at very small values of γ, so that the error parameter E need not be chosen with as much care as when using only the maximum as the envelope. We use the known Taylor series of $g_{\boldsymbol{b}}(\gamma)$ developed at $\gamma = 0$, which is given by (Barouch and Kaufman (1976a), p. 13; Nair and Wang (1989), p. 430)[96]

$$g_{\boldsymbol{b}}(\gamma) = \prod_{j=1}^{k} b_j \frac{\gamma^{k-1}}{(k-1)!} + O(\gamma^k). \tag{B.13}$$

Since a different A needs to be computed for each γ, we evaluate (B.10) iteratively at equally-spaced grid points. The iterative evaluation has the following advantage: if we know that $g_{\boldsymbol{b}}(\gamma - \Delta\gamma) < g_{\boldsymbol{b}}(\gamma)$, where $\Delta\gamma$ is the distance between adjacent grid points, it is possible to use the following inequality

$$\frac{1}{g_{\boldsymbol{b}}(\gamma)} < \frac{1}{g_{\boldsymbol{b}}(\gamma - \Delta\gamma)} \approx \frac{1}{\widetilde{g}_{\boldsymbol{b}}(\gamma - \Delta\gamma)}, \tag{B.14}$$

where the approximation is valid provided that $\widetilde{g}_{\boldsymbol{b}}(\gamma - \Delta\gamma)$ has been calculated without large error.

We propose to construct the envelope for $g_{\boldsymbol{b}}(\gamma)$ from the minimum of the Taylor approximation (i.e. (B.13) without the $O(\gamma^k)$-terms) and the unique maximum of the density. Denoting the unique maximum by $\overline{g}$ and also using (B.14), we can bound the relative discretization error of the approximation (B.10) as follows:

$$\frac{\widetilde{g}_{\boldsymbol{b}}(\gamma) - g_{\boldsymbol{b}}(\gamma)}{g_{\boldsymbol{b}}(\gamma)} = \sum_{m=1}^{\infty} \mathrm{e}^{-mA} \frac{g_{\boldsymbol{b}}(\gamma(2m+1))}{g_{\boldsymbol{b}}(\gamma)} \leq \sum_{m=1}^{\infty} \mathrm{e}^{-mA} \min\Big\{(2m+1)^{k-1}, \frac{\overline{g}}{\widetilde{g}_{\boldsymbol{b}}(\gamma - \Delta\gamma)}\Big\}.$$

Letting M^* be the last index where $(2m+1)^{k-1}$ is the minimum, and setting $M^* = 0$ if this is never the case, we get that

$$M^* = \max\Big\{\Big\lfloor -\frac{1}{2} + \frac{1}{2}\Big(\frac{\overline{g}}{\widetilde{g}_{\boldsymbol{b}}(\gamma - \Delta\gamma)}\Big)^{\frac{1}{k-1}}\Big\rfloor, 0\Big\},$$

where $\lfloor\cdot\rfloor$ is the floor function, and that

$$\frac{\widetilde{g}_{\boldsymbol{b}}(\gamma) - g_{\boldsymbol{b}}(\gamma)}{g_{\boldsymbol{b}}(\gamma)} \leq \begin{cases} \Big(\sum_{m=1}^{M^*} (2m+1)^{k-1} \mathrm{e}^{-mA}\Big) + \frac{\overline{g}}{\widetilde{g}_{\boldsymbol{b}}(\gamma-\Delta\gamma)} \mathrm{e}^{-(M^*+1)A}, & \text{if } M^* \geq 1, \\ \frac{\overline{g}}{\widetilde{g}_{\boldsymbol{b}}(\gamma-\Delta\gamma)} \mathrm{e}^{-A}, & \text{if } M^* = 0. \end{cases} \tag{B.15}$$

[93] Abate and Whitt (1995) (p. 38) propose to use $\Delta M = 11$. In our applications we always use $\Delta M = 11$, and use $M = 100$, increasing M to 500 for very small values of γ.

[94] We use a five-digit accuracy ($E = 5$), which worked well for all values of γ.

[95] For example one may compute a kernel estimate of the density at its median, which usually is not far away from the mode since $g_{\boldsymbol{b}}(\gamma)$ is sufficiently symmetric already for modest values of k.

[96] To see this, evaluate the first $k-1$ derivatives of equation (5c) in McGill and Gibbon (1965) at zero.

Here we have omitted the factor $1/(1-\exp(-A))$ from summing the geometric series because $\exp(-A)$ is negligible. For $M^*=0$, the right-hand side of (B.15) is easily solved for A to achieve a relative error of 10^{-E}. For $M^*\geq 1$, we need a root-finding algorithm to determine A.[97] This requires feeding lower and upper bounds $A_{\min}$ and $A_{\max}$ into the root-finding algorithm, so that the right-hand side of (B.15) is greater than 10^{-E} at $A_{\min}$ and smaller than 10^{-E} at $A_{\max}$.

We are left with deriving suitable values for $A_{\min}$ and $A_{\max}$. Starting with $A_{\min}$, the right-hand side of (B.15) is obviously larger than its first summand, $\exp((k-1)\ln 3 - A_{\min})$, which is greater or equal to 1 already if $A_{\min}\leq (k-1)\ln 3\approx k-1$. For this reason one can use $A_{\min}=k-1$. Continuing with $A_{\max}$, note that for any $A_{\max}>(k-1)\ln 3$ the following inequalities hold

$$0>(k-1)\ln(2\cdot 1+1)-1A_{\max}>(k-1)\ln(2\cdot 2+1)-2A_{\max}>\ldots>(k-1)\ln(2\cdot m+1)-mA_{\max}.$$

Thus, the sum over $m=1,\ldots,M^*$ in (B.15) is always less than $M^*\exp((k-1)\ln 3-A_{\max})$, so that the right-hand side of (B.15) is always less than

$$2\cdot\max\Big\{M^*\exp((k-1)\ln 3-A_{\max}),\tfrac{\overline{g}}{\widetilde{g}_{\boldsymbol{b}}(\gamma-\Delta\gamma)}\exp(-(1+M^*)A_{\max})\Big\}.$$

Hence, we know that the right-hand side of (B.15) remains below 10^{-E} if we choose $A_{\max}$ as

$$A_{\max}=\max\Big\{\ln(2M^*)+(k-1)\ln 3+E\ln 10,\frac{1}{1+M^*}\big(\ln(2\overline{g})-\ln\widetilde{g}_{\boldsymbol{b}}(\gamma-\Delta\gamma)+E\ln 10\big)\Big\}.$$

B.3 Maximum Likelihood Estimation

In this section we again show the parameter dependence and separate the truncation factor in the field-size density as $f_{\boldsymbol{\theta}}/p_{\mathrm{tr}}(\boldsymbol{\theta})$.

Point estimation

At first, we state the important relation that for any distribution from the exponential family connects the log-derivatives of the parameter dependent function $a(\boldsymbol{\eta})$ to the mean vector and covariance matrix of the sufficient statistics $\boldsymbol{s}(x)$. If X is distributed with pdf $c(x)\mathrm{e}^{\boldsymbol{\eta}^\top\boldsymbol{s}(x)}/a(\boldsymbol{\eta})$, then it holds that (see Dempster et al. (1977), p. 5)

$$\begin{aligned} a(\boldsymbol{\eta}) &= \int c(x)\mathrm{e}^{\boldsymbol{\eta}^\top\boldsymbol{s}(x)}\mathrm{d}x,\\ \Rightarrow \frac{\partial \ln a(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} &= \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(X)\big],\\ \Rightarrow \frac{\partial^2 \ln a(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top} &= \mathbb{V}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(X)\big]. \end{aligned} \tag{B.16}$$

By truncating a pdf from the exponential family, the only change in the exponential family form is that the parameter dependent function needs to be multiplied by the truncation

[97] We use the R-method "uniroot". We implement this by solving the natural logarithm of (B.15) for $-E\ln 10$.

factor. Thus, the pdf of $\tilde{Y}_1$ has the parameter dependent function $a(\boldsymbol{\eta})(1 - p_{\text{tr}}(\boldsymbol{\eta}))$. Moreover, the density of the remaining field sizes in equation (4.15) can be written in exponential family form as

$$\rho_{\boldsymbol{\eta}}(x|\gamma) = \big[\mathrm{e}^{-\gamma x^{\beta}} c(x)\big] \mathrm{e}^{\boldsymbol{\eta}^{\top} \boldsymbol{s}(x)} / \big[a(\boldsymbol{\eta}) \phi_{\boldsymbol{\eta}}(\gamma) p_{\text{tr}}(\boldsymbol{\eta})\big], \tag{B.17}$$

showing that it has the parameter dependent function $a(\boldsymbol{\eta})\phi_{\boldsymbol{\eta}}(\gamma)p_{\text{tr}}(\boldsymbol{\eta})$. Table B1 summarizes several important formulas for the exponential family form of the lognormal distribution.[98]

Table B1: Exponential family form of the lognormal distribution

$\big(\eta_1(\mu,\sigma), \eta_2(\mu,\sigma)\big) = \big(\mu\sigma^{-2}, (-2\sigma^2)^{-1}\big)$ $\big(\mu(\eta_1,\eta_2), \sigma(\eta_1,\eta_2)\big) = \big(\eta_1(-2\eta_2)^{-1}, (-2\eta_2)^{-\frac{1}{2}}\big)$		
$c(x) = \frac{1}{\sqrt{2\pi}x},\ s_1(x) = \ln x,\ s_2(x) = (\ln x)^2,\ \ln a(\boldsymbol{\eta}) = -\frac{1}{4}\frac{\eta_1^2}{\eta_2} - \frac{1}{2}\ln(-2\eta_2)$		
Jacobian of $\boldsymbol{\eta} \to \boldsymbol{\theta}$:	$\mu(\eta_1,\eta_2)$	$\sigma(\eta_1,\eta_2)$
Derivative w.r.t. η_1	σ^2	0
Derivative w.r.t. η_2	$2\mu\sigma^2$	σ^3
Gradient of $\ln a(\boldsymbol{\eta})$:	$\partial \ln a(\boldsymbol{\theta})/\partial\eta_1 = \mu,$	$\partial \ln a(\boldsymbol{\theta})/\partial\eta_2 = \mu^2 + \sigma^2$
Hessian of $\ln a(\boldsymbol{\eta})$:	$\partial \ln a(\boldsymbol{\theta})/\partial\eta_1$	$\partial \ln a(\boldsymbol{\theta})/\partial\eta_2$
Derivative w.r.t. η_1	σ^2	$2\mu\sigma^2$
Derivative w.r.t. η_2	$2\mu\sigma^2$	$4\mu^2\sigma^2 + 2\sigma^4$

We continue from equation (4.17) by reparameterizing $\boldsymbol{\theta} \to \boldsymbol{\eta}$ so that the exponential family form of the function becomes

$$\begin{aligned} F(\boldsymbol{\eta}', \nu'|\boldsymbol{\eta}, \nu) = k \cdot \Big(\frac{1}{k}\sum_{i=1}^{k} \boldsymbol{s}(x_i)^{\top}\boldsymbol{\eta}'\Big) - k\ln a(\boldsymbol{\eta}') + \mathbb{E}_{(\boldsymbol{\eta},\nu)}\Big\{\Big(\sum_{i=1}^{M} \boldsymbol{s}(\tilde{Y}_i)^{\top}\boldsymbol{\eta}'\Big) - M\ln a(\boldsymbol{\eta}')\Big|\mathcal{F}\Big\} \\ + \mathbb{E}_{(\boldsymbol{\eta},\nu)}\Big\{\Big(\sum_{i=k+1}^{N} \boldsymbol{s}(\tilde{X}_i)^{\top}\boldsymbol{\eta}'\Big) - (N-k)\ln a(\boldsymbol{\eta}')\Big|\mathcal{F}\Big\} \\ - \nu' + \ln(\nu')\mathbb{E}_{(\boldsymbol{\eta},\nu)}\big[N\big|\mathcal{F}\big] + \text{constant}. \end{aligned}$$

[98] For the Pareto distribution the exponential family form is derived as follows. Rearrange the density as $f_\theta(x) = \theta \underline{x}^{\theta}/x^{\theta+1} = (1/x)e^{-\theta(\ln x - \ln \underline{x})}(1/\theta^{-1})$. Thus, $\eta(\theta) = -\theta$, $s(x) = \ln x - \ln \underline{x}$, and $\ln a(\eta) = -\ln(-\eta)$. Taking derivatives with respect to η and plugging in $\eta(\theta)$ yields $\partial \ln a(\theta)/\partial\eta = 1/\theta$, $\partial^2 \ln a(\theta)/\partial\eta^2 = 1/\theta^2$.

Thus, the necessary conditions for maximization with respect to $(\boldsymbol{\eta}', \nu')$ are

$$
\begin{aligned}
0 &\overset{!}{=} \frac{\partial F(\boldsymbol{\eta}', \nu' | \boldsymbol{\eta}, \nu)}{\partial \boldsymbol{\eta}'} \\
\Leftrightarrow \mathbb{E}_{(\boldsymbol{\eta},\nu)}[N + M | \mathcal{F}] \frac{\partial \ln a(\boldsymbol{\eta}')}{\partial \boldsymbol{\eta}'} &= k \cdot \Big(\frac{1}{k} \sum_{i=1}^{k} \boldsymbol{s}(x_i) \Big) + \mathbb{E}_{(\boldsymbol{\eta},\nu)}[M | \mathcal{F}] \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{Y}_1)\big] \\
&\quad + \mathbb{E}_{(\boldsymbol{\eta},\nu)}\Big\{ \sum_{i=k+1}^{N} \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{X}_i) \big| N, \Gamma_k, \mathcal{F}\big] \Big| \mathcal{F} \Big\} \\
\Leftrightarrow \frac{\mathbb{E}_{(\boldsymbol{\eta},\nu)}[N | \mathcal{F}]}{p_{\mathrm{tr}}(\boldsymbol{\eta})} \frac{\partial \ln a(\boldsymbol{\eta}')}{\partial \boldsymbol{\eta}'} &= k \cdot \Big(\frac{1}{k} \sum_{i=1}^{k} \boldsymbol{s}(x_i) \Big) + \mathbb{E}_{(\boldsymbol{\eta},\nu)}[N | \mathcal{F}] \frac{1 - p_{\mathrm{tr}}(\boldsymbol{\eta})}{p_{\mathrm{tr}}(\boldsymbol{\eta})} \cdot \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{Y}_1)\big] \\
&\quad + \mathbb{E}_{(\boldsymbol{\eta},\nu)}\Big\{ (N - k) \cdot \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{X}_{k+1}) \big| \Gamma_k, \mathcal{F}\big] \Big| \mathcal{F} \Big\} \\
\Leftrightarrow \frac{\partial \ln a(\boldsymbol{\eta}')}{\partial \boldsymbol{\eta}'} &= p_{\mathrm{tr}}(\boldsymbol{\eta}) \alpha(\boldsymbol{\eta}, \nu) \Big(\frac{1}{k} \sum_{i=1}^{k} \boldsymbol{s}(x_i) \Big) + \big(1 - p_{\mathrm{tr}}(\boldsymbol{\eta})\big) \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{Y}_1)\big] \\
&\quad + p_{\mathrm{tr}}(\boldsymbol{\eta}) \big(1 - \alpha(\boldsymbol{\eta}, \nu)\big) \int_0^\infty \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{X}_{k+1}) \big| \Gamma_k = \gamma, \mathcal{F}\big] g^{(1)}_{(\boldsymbol{\eta},\nu)}(\gamma | \mathcal{F}) \mathrm{d}\gamma,
\end{aligned}
$$

where $\alpha(\boldsymbol{\eta}, \nu) = k / \mathbb{E}_{(\boldsymbol{\eta},\nu)}[N \big| \mathcal{F}]$, and

$$
\begin{aligned}
0 &\overset{!}{=} \frac{\partial F(\boldsymbol{\eta}', \nu' | \boldsymbol{\eta}, \nu)}{\partial \nu'}, \\
\Leftrightarrow \nu' &= \mathbb{E}_{(\boldsymbol{\eta},\nu)}[N | \mathcal{F}].
\end{aligned}
$$

In the above derivations, we proceed after taking the derivative by evaluating the expectations stepwise with the law of iterated expectations. To arrive at the third equality, we first use the fact that the Negbin(n, p)-distribution has the mean $n(1-p)/p$ which implies that

$$
\begin{aligned}
\mathbb{E}_{(\boldsymbol{\eta},\nu)}[M | \mathcal{F}] &= \mathbb{E}_{(\boldsymbol{\eta},\nu)}[N | \mathcal{F}](1 - p_{\mathrm{tr}}(\boldsymbol{\eta}))/p_{\mathrm{tr}}(\boldsymbol{\eta}), \\
\mathbb{E}_{(\boldsymbol{\eta},\nu)}[N + M | \mathcal{F}] &= \mathbb{E}_{(\boldsymbol{\eta},\nu)}[N | \mathcal{F}]/p_{\mathrm{tr}}(\boldsymbol{\eta}).
\end{aligned}
$$

To arrive the third equality, we also use the fact that

$$
\mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{X}_i) \big| N, \Gamma_k, \mathcal{F}\big] = \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{X}_i) \big| \Gamma_k, \mathcal{F}\big] = \mathbb{E}_{\boldsymbol{\eta}}\big[\boldsymbol{s}(\tilde{X}_{k+1}) \big| \Gamma_k, \mathcal{F}\big]
$$

for all $i = k+1, \ldots, N$, as can be seen from equation (4.14). To arrive at the fourth equality, we apply Lemma 4 (eq. B.7) to the outer expected value over the joint distribution of $(N, \Gamma_k) | \mathcal{F}$.

Finally, note that the necessary conditions are sufficient for $(\boldsymbol{\eta}', \nu')$ to be a maximum given $(\boldsymbol{\eta}, \nu)$. The function $F(\cdot | \cdot)$ is globally concave in $(\boldsymbol{\eta}', \nu')$ since its hessian with respect to $(\boldsymbol{\eta}', \nu')$ equals

$$
\frac{\partial^2 F(\boldsymbol{\eta}', \nu' | \boldsymbol{\eta}, \nu)}{\partial (\boldsymbol{\eta}', \nu') \partial (\boldsymbol{\eta}', \nu')^\top} = \begin{bmatrix} -\frac{\mathbb{E}_{(\boldsymbol{\eta},\nu)}[N | \mathcal{F}]}{p_{\mathrm{tr}}(\boldsymbol{\eta})} \frac{\partial^2 \ln a(\boldsymbol{\eta}')}{\partial \boldsymbol{\eta}' \boldsymbol{\eta}'^\top} & 0 \\ 0 & -\frac{\mathbb{E}_{(\boldsymbol{\eta},\nu)}[N | \mathcal{F}]}{(\nu')^2} \end{bmatrix},
$$

where $\frac{\partial^2 \ln a(\boldsymbol{\eta}')}{\partial \boldsymbol{\eta}' \boldsymbol{\eta}'^\top}$ is positive definite for each $\boldsymbol{\eta}'$, as can be seen by the relation to a covariance matrix shown in (B.16). Thus, independent of $(\boldsymbol{\eta}', \nu')$ all eigenvalues of the hessian are negative, which implies that it is globally negative definite.

Interval estimation

By reparametrizing equation (4.13) into exponential family form, and taking second-order derivatives, we arrive at

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\theta},\nu)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top} =& -k\Big(\frac{\partial^2 \ln a(\boldsymbol{\theta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}+\frac{\partial^2 \ln p_{\mathrm{tr}}(\boldsymbol{\theta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}\Big)\\ &+\Big\{\int_0^\infty \frac{\partial^2\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}\frac{1}{\phi_{\boldsymbol{\theta}}(\gamma)}\nu\phi_{\boldsymbol{\theta}}(\gamma)\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\\ &+\int_0^\infty \frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}}\frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}^\top}\big(\nu\phi_{\boldsymbol{\theta}}(\gamma)\big)^2\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\\ &-\Big(\int_0^\infty \frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}}\nu\phi_{\boldsymbol{\theta}}(\gamma)\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\Big)\\ &\cdot\Big(\int_0^\infty \frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}^\top}\nu\phi_{\boldsymbol{\theta}}(\gamma)\mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\Big)\Big\}\Big/\Big(\int_0^\infty \mathrm{e}^{\nu(\phi_{\boldsymbol{\theta}}(\gamma)-1)}g_{\boldsymbol{b}}(\gamma)\mathrm{d}\gamma\Big).\end{aligned}$$

Using $\frac{\partial^2\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}\frac{1}{\phi_{\boldsymbol{\theta}}(\gamma)} = \frac{\partial^2 \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}+\frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}}\frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}^\top}$ to split the term in the second line, then combining the resulting $\frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}}\frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}^\top}$-term with the term in the third line, and finally using the formulas from (B.5-B.6) yields

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\theta},\nu)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top} =& -k\Big(\frac{\partial^2 \ln a(\boldsymbol{\theta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}+\frac{\partial^2 \ln p_{\mathrm{tr}}(\boldsymbol{\theta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}\Big)\\ &+\mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}]\int_0^\infty \frac{\partial^2 \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\\ &+\mathbb{E}_{(\boldsymbol{\theta},\nu)}[(N-k)^2|\mathcal{F}]\Big\{\int_0^\infty \frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}}\frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}^\top}g^{(2)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\}\\ &-\mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}]^2\Big\{\int_0^\infty \frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}}g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\}\Big\{\int_0^\infty \frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}^\top}g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\}.\end{aligned}$$

Moreover, the relations from (B.16-B.17) imply

$$\begin{aligned}\frac{\partial^2 \ln a(\boldsymbol{\theta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}+\frac{\partial^2 \ln p_{\mathrm{tr}}(\boldsymbol{\theta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top} &= \mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big],\\ \frac{\partial^2 \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top} = \frac{\partial^2 \ln[a(\boldsymbol{\theta})p_{\mathrm{tr}}(\boldsymbol{\theta})\phi_{\boldsymbol{\theta}}(\gamma)]}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top}-\frac{\partial^2 \ln[a(\boldsymbol{\theta})p_{\mathrm{tr}}(\boldsymbol{\theta})]}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top} &= \mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big],\\ \frac{\partial \ln\phi_{\boldsymbol{\theta}}(\gamma)}{\partial\boldsymbol{\eta}} = \frac{\partial \ln[a(\boldsymbol{\theta})p_{\mathrm{tr}}(\boldsymbol{\theta})\phi_{\boldsymbol{\theta}}(\gamma)]}{\partial\boldsymbol{\eta}}-\frac{\partial \ln[a(\boldsymbol{\theta})p_{\mathrm{tr}}(\boldsymbol{\theta})]}{\partial\boldsymbol{\eta}} &= \mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big],\end{aligned}$$

where $\tilde{X}_{k+1}|\Gamma_k$ is distributed with pdf $\rho_{\boldsymbol{\theta}}(x|\gamma)$ and $\tilde{X}_1$ is distributed with pdf $f_{\boldsymbol{\theta}}(x)/p_{\mathrm{tr}}(\boldsymbol{\theta})$.

This finally results in the following formula for the Hessian of the log-likelihood:

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\theta},\nu)}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\top} =& -k\cdot\mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\\ &+\ \mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}]\cdot\Big\{\int_0^\infty\Big(\mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{V}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\}\\ &+\ \mathbb{E}_{(\boldsymbol{\theta},\nu)}[(N-k)^2|\mathcal{F}]\cdot\Big\{\int_0^\infty\Big(\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)\\ &\qquad\cdot\Big(\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)^\top g^{(2)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\}\\ &-\ (\mathbb{E}_{(\boldsymbol{\theta},\nu)}[N-k|\mathcal{F}])^2\cdot\Big\{\int_0^\infty\Big(\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\}\\ &\qquad\cdot\Big\{\int_0^\infty\Big(\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_{k+1})\big|\Gamma_k=\gamma,\mathcal{F}\big]-\mathbb{E}_{\boldsymbol{\theta}}\big[\boldsymbol{s}(\tilde{X}_1)\big]\Big)^\top g^{(1)}_{(\boldsymbol{\theta},\nu)}(\gamma|\mathcal{F})\mathrm{d}\gamma\Big\}.\end{aligned}$$

B.4 Proof of Proposition 4a and 4b

Proof of Proposition 4a: When transition rates are constant at $R_i = C$ for all $i = k+1, k+2, \ldots$, the process $(M_u)_{u\geq 0}$ is a Poisson process with intensity parameter C and starting value $M_0 = k$, that is $(M_u - k) \sim \text{Pois}(Cu)$, or

$$\mathbb{P}[M_u - k = l] = \mathrm{e}^{-Cu}\frac{(Cu)^l}{l!} \text{ for each } u > 0. \tag{B.18}$$

Without loss of generality assume $\lim_{t\to\infty}\Lambda_{\mathrm{a}}(t) = 1$ and $C = \nu U_k$. This implies

$$\lim_{t\to\infty} N_t = \lim_{u\uparrow 1} M_u = M_1,$$

where the first equality follows from $\Lambda_{\mathrm{a}}(t) \leq 1$ and $\lim_{t\to\infty}\Lambda_{\mathrm{a}}(t) = 1$, and the second equality holds with probability one (because there can be a jump exactly at $u = 1$, but this event has probability zero).

Thus, plugging $u = 1$ and $C = \nu U_k$ into (B.18) establishes that

$$\mathbb{P}\big[\big(\lim_{t\to\infty} N_t\big) - k = l\big|U_k\big] = \mathbb{P}[M_1 - k = l|U_k] = \mathrm{e}^{-\nu U_k}\frac{(\nu U_k)^l}{l!}.$$

□

Proof of Proposition 4b: We condition on a particular value for $\mathcal{N}$, i.e. let $\mathcal{N} = n \geq k$. The requirement $R_i = 0$ for $i \geq n+1$ implies that $\mathbb{P}[M_z = n] = \mathbb{P}[M_z \geq n]$. The next step uses the equivalence $\mathbb{P}[M_z \geq n] = \mathbb{P}[Z_n \leq z]$, which is a fundamental property of counting processes. Because $\mathbb{P}[Z_n \leq z]$ is the cdf of a random variable supported on the nonnegative real numbers (Z_n is the sum of $n-k$ of exponentially distributed variables), it follows that

$$\lim_{z\to\infty}\mathbb{P}[M_z = n] = \lim_{z\to\infty}\mathbb{P}[Z_n \leq z] = 1.$$

Using $N_t = M_{\Lambda_b(t)}$ shows that this carries over to N_t if the function $\Lambda_b(t)$ is unbounded, yielding $\lim_{t\to\infty} \mathbb{P}[N_t = n] = 1$. Removing the conditioning on $\mathcal{N} = n$, this clearly implies that

$$\lim_{t\to\infty} \mathbb{P}[|N_t - \mathcal{N}| \leq \epsilon] = 1, \text{ for each } \epsilon > 0,$$

which means that N_t converges in probability to $\mathcal{N}$. As convergence in probability implies convergence in distribution, the proof is complete. □

B.5 Further Derivations for the Size-Biased Sampling Model

We now prepare the derivation of the induced order statistics result for $(X_1, \ldots, X_n)$ with the following Lemma. To give a concise notation for order statistics we need the following definitions. Let $\Delta_n(c)$ be the unordered and $\Delta_n^*(a, b)$ be the ordered n-dimensional, open simplex over the intervals $(0, c)$ and (a, b), respectively, i.e.

$$\begin{aligned}\Delta_n(c) =& \big\{(\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n \big| \epsilon_1, \ldots, \epsilon_n > 0, \ \sum_{i=1}^n \epsilon_i < c\big\},\\ \Delta_n^*(a, b) =& \big\{(\gamma_1, \ldots, \gamma_n) \in \mathbb{R}^n \big| a < \gamma_1 < \gamma_2 < \ldots < \gamma_n < b\big\}.\end{aligned}$$

Note that with $c = b-a$, the latter set is a coordinate transform of the former set resulting as the partial sums plus a shift by a.

Lemma 5. *Let $w_1, \ldots, w_k > 0$ and $s \geq 0$. Upon defining the partial sums $b_i = w_i + \ldots + w_k$ for each $i = 1, \ldots, k$, the term $\prod_{i=1}^k (b_i + s)^{-1}$ has the two integral representations*

$$\begin{aligned}\prod_{i=1}^k (b_i + s)^{-1} &= \int_0^\infty \int_{\gamma_1}^\infty \cdots \int_{\gamma_{k-1}}^\infty \mathrm{e}^{-\gamma_k s} \prod_{i=1}^k \mathrm{e}^{-\gamma_i w_i} \mathrm{d}\gamma_i,\\ \prod_{i=1}^k (b_i + s)^{-1} &= \prod_{i=1}^k b_i^{-1} \int_0^\infty \mathrm{e}^{-\gamma_k s} g_{\boldsymbol{b}}(\gamma_k) \mathrm{d}\gamma_k,\end{aligned} \tag{B.19}$$

where $g_{\boldsymbol{b}}(\gamma_k)$ is the general gamma density with parameters $\boldsymbol{b} = (b_1, \ldots, b_k)$.

Proof of Lemma 5:

First, recall the straight-forward property of the exponential function that $\int_0^\infty \mathrm{e}^{-\epsilon(b+s)} \mathrm{d}\epsilon = (b + s)^{-1}$. Upon applying this for all $i = 1, \ldots, k$ we get

$$\begin{aligned}\prod_{i=1}^k (b_i + s)^{-1} &= \prod_{i=1}^k \Big[\int_0^\infty \mathrm{e}^{-\epsilon_i (b_i + s)} \mathrm{d}\epsilon_i \Big]\\ &= \int_0^\infty \cdots \int_0^\infty \mathrm{e}^{-(\epsilon_1 + \ldots + \epsilon_k)s} \mathrm{e}^{-(\epsilon_1 b_1 + \ldots + \epsilon_k b_k)} \mathrm{d}\epsilon_1 \cdots \mathrm{d}\epsilon_k.\end{aligned} \tag{B.20}$$

Here the product over the integrals can be expanded just like a product over finite sums since all integrands are positive, allowing arbitrary interchange of the order of integration by Fubini's theorem.

To proceed, we rearrange the second exponent from (B.20) by applying a version of Abel's summation by parts, which is the discrete analogue to integration by parts as applicable

to sequences. Upon defining $\gamma_i = \epsilon_1 + \ldots + \epsilon_i$ for each $i = 1, \ldots, k$, the summation by parts formula is expressible as (see, e.g., Königsberger (2003), p. 305)[99]

$$\begin{aligned}
&\epsilon_1(w_1 + \ldots + w_k) + \epsilon_2(w_2 + \ldots + w_k) + \ldots + \epsilon_k w_k \\
&= \gamma_1 b_1 + (\gamma_2 - \gamma_1)b_2 + \ldots + (\gamma_k - \gamma_{k-1})b_k \\
&= \gamma_1(b_1 - b_2) + \ldots + \gamma_{k-1}(b_{k-1} - b_k) + \gamma_k b_k \\
&= \epsilon_1 w_1 + \ldots + (\epsilon_1 + \ldots + \epsilon_{k-1})w_{k-1} + (\epsilon_1 + \ldots + \epsilon_k)w_k.
\end{aligned} \tag{B.21}$$

Then, the multivariate substitution $(\epsilon_1, \ldots, \epsilon_k) \to (\gamma_1, \ldots, \gamma_k)$ yields via the substitution rule for multivariate mappings[100]

$$\begin{aligned}
\prod_{i=1}^{k}(b_i + s)^{-1} &= \int_0^\infty \cdots \int_0^\infty \mathrm{e}^{-(\epsilon_1 + \ldots + \epsilon_k)s} \mathrm{e}^{-\{\epsilon_1 w_1 + \ldots + (\epsilon_1 + \ldots + \epsilon_k)w_k\}} \mathrm{d}\epsilon_1 \cdots \mathrm{d}\epsilon_k \\
&= \int_0^\infty \int_{\gamma_1}^\infty \cdots \int_{\gamma_{k-1}}^\infty \mathrm{e}^{-\gamma_k s} \mathrm{e}^{-(\gamma_1 w_1 + \ldots + \gamma_k w_k)} \mathrm{d}\gamma_1 \cdots \mathrm{d}\gamma_k.
\end{aligned}$$

To prove the second part, we proceed from (B.20) directly by substituting $(\epsilon_1, \ldots, \epsilon_{k-1}, \epsilon_k) \to (\epsilon_1, \ldots, \epsilon_{k-1}, \gamma_k)$, i.e. we substitute only $\gamma_k = \epsilon_1 + \ldots + \epsilon_k$:

$$\begin{aligned}
\prod_{i=1}^{k}(b_i + s)^{-1} &= \int_0^\infty \int_0^\infty \cdots \int_0^\infty \mathbb{1}(\epsilon_1 + \ldots + \epsilon_{k-1} < \gamma_k) \mathrm{e}^{-\gamma_k s} \mathrm{e}^{-\{\epsilon_1 b_1 + \ldots + (\gamma_k - \epsilon_1 - \ldots - \epsilon_{k-1})b_k\}} \mathrm{d}\epsilon_1 \cdots \mathrm{d}\epsilon_{k-1} \mathrm{d}\gamma_k \\
&= \prod_{i=1}^{k} b_i^{-1} \int_0^\infty \mathrm{e}^{-\gamma_k s} \Bigg\{ \int_{\Delta_{k-1}(\gamma_k)} \Big(\prod_{i=1}^{k-1} b_i \mathrm{e}^{-\epsilon_i b_i} \Big) b_k \mathrm{e}^{-(\gamma_k - \epsilon_1 - \ldots - \epsilon_{k-1})b_k} \mathrm{d}\epsilon_1 \cdots \mathrm{d}\epsilon_{k-1} \Bigg\} \mathrm{d}\gamma_k.
\end{aligned}$$

The proof is completed by noting that the integral in the curly brackets is precisely the convolution of k exponential densities with parameters $b_1, \ldots, b_k$. □

Both representations from Lemma 5 have been used in the size-biased sampling literature for further derivations. Starting with the second representation, multiplying by $\prod_{i=1}^{k} b_i$ and setting $s = \tilde{x}_{k+1}^{\beta} + \ldots + \tilde{x}_n^{\beta}$ proves the first line of equation (4.12).

The first representation allows to derive the induced order statistics result, which follows in the next subsection.

[99] For this case of finite-dimensional vectors, the summation by parts formula is also easily derived by simple matrix algebra: First, write the vector $(b_1, \ldots, b_k)^\top$ as the product of the upper triangular matrix filled with 1s and $(w_1, \ldots, w_k)^\top$. Then, rearrange the sum $\epsilon_1 b_1 + \ldots + \epsilon_k b_k$ as

$$\begin{aligned}
&(\epsilon_1, \ldots, \epsilon_k) \cdot (w_1 + \ldots + w_k, \ldots, w_k)^\top \\
&= (\epsilon_1, \ldots, \epsilon_k) \cdot \left(\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix} \right) = \left(\begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_k \end{pmatrix} \right)^\top \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix} \\
&= (\epsilon_1, \ldots, \epsilon_1 + \ldots + \epsilon_k) \cdot (w_1, \ldots, w_k)^\top.
\end{aligned}$$

[100] Note from the last footnote that the Jacobian of this mapping is the lower triangular matrix of 1s, so that the Jacobian and its inverse have a constant determinant of 1.

A priori: the joint distribution of all field sizes in the order of discovery

We use the first integral representation from Lemma 5 with $w_i = x_i^\beta$, $i = 1, \ldots, n$, and $s = 0$ to rearrange equation (4.7). Factoring all terms inside the integrals, we get

$$\begin{aligned}\mathbb{P}[X_1 \in \mathrm{d}x_1, \ldots, X_n \in \mathrm{d}x_n | N = n] &= n! \prod_{i=1}^{n} \frac{x_i^\beta}{x_i^\beta + \ldots + x_n^\beta} f(x_i)\mathrm{d}x_i \\ &= \int_0^\infty \int_{\gamma_1}^\infty \cdots \int_{\gamma_{n-1}}^\infty n! \prod_{i=1}^{n} x_i^\beta \mathrm{e}^{-\gamma_i x_i^\beta} f(x_i)\mathrm{d}x_i \mathrm{d}\gamma_i.\end{aligned} \tag{B.22}$$

Following Pitman and Tran (2015) (pp. 2489ff.), (B.22) can be interpreted as follows: let $\tilde{X}_1, \ldots, \tilde{X}_n$ be iid as in assumption (1), $\varepsilon_1, \ldots, \varepsilon_n$ be iid standard exponential rvs, where the ε_i are also independent of the $\tilde{X}_i$, and define $\tilde{\Gamma}_i = \varepsilon_i / \tilde{X}_i^\beta$. By the fact that $\big((\tilde{X}_i, \tilde{\Gamma}_i), i = 1, \ldots, n\big)$ form n iid pairs from the joint pdf $x^\beta \mathrm{e}^{-\gamma x^\beta} f(x)$, and noting that $n! = |\mathcal{I}_n|$, we can restate (B.22) in probabilistic terms as

$$\begin{aligned}&\mathbb{P}[X_1 \in \mathrm{d}x_1, \ldots, X_n \in \mathrm{d}x_n | N = n] \\ &= \int_0^\infty \int_{\gamma_1}^\infty \cdots \int_{\gamma_{n-1}}^\infty \sum_{(i(1),\ldots,i(n)) \in \mathcal{I}_n} \mathbb{P}\big[\tilde{\Gamma}_{i(1)} \in \mathrm{d}\gamma_1, \tilde{X}_{i(1)} \in \mathrm{d}x_1, \ldots, \tilde{\Gamma}_{i(n)} \in \mathrm{d}\gamma_n, \tilde{X}_{i(n)} \in \mathrm{d}x_n\big] \\ &= \mathbb{P}\Big[\bigcup_{(i(1),\ldots,i(n)) \in \mathcal{I}_n} \{\tilde{\Gamma}_{i(1)} < \ldots < \tilde{\Gamma}_{i(n)}, \tilde{X}_{i(1)} \in \mathrm{d}x_1, \ldots, \tilde{X}_{i(n)} \in \mathrm{d}x_n\}\Big].\end{aligned} \tag{B.23}$$

For a fixed set of size-measures, $\{x_1, \ldots, x_n\}$, this interpretation was first proved by Gordon (1983). (B.23) expresses the probability that the $\tilde{X}_i$ which becomes the first size-biased pick has the smallest $\tilde{\Gamma}_i$, the one which becomes the second size-biased pick has the second smallest $\tilde{\Gamma}_i$, etc., which means precisely that $(X_1, \ldots, X_n)$ are the induced order statistics or concomitants of $(\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_n)$ (Pitman and Tran (2015), p. 2489).

We proceed by expanding in (B.22) each term by (minus one times) the derivative of the Laplace transform.. Due to $-\phi'(\gamma) = \int_{\underline{x}}^{\overline{x}} x^\beta \mathrm{e}^{-\gamma x^\beta} f(x)\mathrm{d}x$, this expansion is mathematically equivalent to refactoring the joint density of $(\tilde{X}_i, \tilde{\Gamma}_i)$ according to Bayes' rule. Then, the joint pdf of $(X_1, \ldots, X_n)$ given $N = n$ can be restated as

$$\int_0^\infty \int_{\gamma_1}^\infty \cdots \int_{\gamma_{n-1}}^\infty \left\{ \prod_{i=1}^{n} \frac{x_i^\beta \mathrm{e}^{-\gamma_i x_i^\beta} f(x_i)\mathrm{d}x_i}{-\phi'(\gamma_i)} \right\} \cdot \left\{ n! \prod_{i=1}^{n} (-\phi'(\gamma_i))\mathrm{d}\gamma_i \right\}. \tag{B.24}$$

We denote the pdf from the first bracket by $h(x_i|\gamma_i) \propto x_i^\beta \mathrm{e}^{-\gamma_i x_i^\beta} f(x_i)$. The second bracket, together with the integral boundaries, shows that $(\Gamma_1, \ldots, \Gamma_n)$ are jointly distributed as (increasing) order statistics from the pdf $-\phi'(\gamma)$ (with cdf $1 - \phi(\gamma)$). Pitman and Tran (2015) (p. 2490) go further to express (B.24) in terms of decreasing uniform order statistics, which immediately follows since $\phi(\gamma)$ is a complementary cdf. Making the corresponding substitution $u_i = \phi(\gamma_i)$, (B.24) becomes

$$\int_0^1 \int_0^{u_1} \cdots \int_0^{u_{n-1}} \left\{ \prod_{i=1}^{n} \frac{x_i^\beta \mathrm{e}^{-\phi^{-1}(u_i) x_i^\beta} f(x_i)\mathrm{d}x_i}{-\phi'(\phi^{-1}(u_i))} \right\} \cdot \left\{ n! \prod_{i=1}^{n} \mathrm{d}u_i \right\}. \tag{B.25}$$

Overall, (B.22-B.25) imply a further procedure with which Monte Carlo draws for $(X_1, \ldots X_n)$ can be obtained: take $(\tilde{X}_1, \ldots, \tilde{X}_n)$ as a random sample from f, then draw $\tilde{\Gamma}_i | \tilde{X}_i$ as $\tilde{\Gamma}_i = \varepsilon_i / \tilde{X}_i^\beta$, and then rearrange the sequence as $\tilde{\Gamma}_{i(1)} < \ldots < \tilde{\Gamma}_{i(n)}$, from which one obtains $X_1 = \tilde{X}_{i(1)}, \ldots, X_n = \tilde{X}_{i(n)}$.

A posteriori: the joint distribution of the remaining field sizes in the order of discovery

We now derive a posterior version of the above equations for the distribution of the remaining field sizes $(X_{k+1}, \ldots, X_N)$ given $(X_1, \ldots, X_k, N)$. In (B.22-B.24), we interchange the order of integration so that the integral over γ_k is the outermost integral, and the integrals over $\gamma_1, \ldots, \gamma_k$ and $\gamma_{k+1}, \ldots, \gamma_n$ are each pooled inside. Then, (B.22-B.24) becomes

$$n! \int_0^\infty \Big\{ p_{(<)}(\boldsymbol{x}; \gamma_k) \mathrm{d}x_1 \cdots \mathrm{d}x_k \Big\} \cdot \int_{\Delta^*_{n-k}(\gamma_k, \infty)} \Big\{ \prod_{i=k+1}^n \frac{x_i^\beta \mathrm{e}^{-\gamma_i x_i^\beta} f(x_i) \mathrm{d}x_i}{-\phi'(\gamma_i)} \Big\} \cdot \Big\{ \prod_{i=k+1}^n (-\phi'(\gamma_i)) \mathrm{d}\gamma_i \Big\} \cdot \mathrm{d}\gamma_k,$$

where we define $p_{(<)}(\boldsymbol{x}; \gamma_k)$ as the density of $(X_1, \ldots, X_k)$ jointly with the probability that the mixing rvs with index below k are smaller than the value γ_k, i.e.

$$p_{(<)}(\boldsymbol{x}; \gamma_k) = \int_{\Delta^*_{k-1}(0, \gamma_k)} \prod_{i=1}^k x_i^\beta \mathrm{e}^{-\gamma_i x_i^\beta} f(x_i) \mathrm{d}\gamma_i.$$

We rearrange the integral over the ordered set $\Delta^*_{k-1}(0, \gamma_k)$ by applying Abel's formula (B.21) with $b_i = x_i^\beta + \ldots + x_k^\beta$, $i = 1, \ldots, k$. That is, we substitute back $(\gamma_1, \ldots, \gamma_{k-1}, \gamma_k) \to (\epsilon_1, \ldots, \epsilon_{k-1}, \gamma_k)$ to rearrange the $p_{(<)}$-term as in the second integral representation of Lemma 5 (with $s = 0$ and without the integral over γ_k which is kept fixed here):

$$\begin{aligned} p_{(<)}(\boldsymbol{x}; \gamma_k) &= \Big\{ \prod_{i=1}^k x_i^\beta f(x_i) \Big\} \cdot \Big\{ \int_{\Delta^*_{k-1}(0, \gamma_k)} \mathrm{e}^{-(\gamma_1 x_1^\beta + \ldots + \gamma_k x_k^\beta)} \mathrm{d}\gamma_1 \ldots \mathrm{d}\gamma_{k-1} \Big\} \\ &= \Big\{ \prod_{i=1}^k \frac{x_i^\beta}{b_i} f(x_i) \Big\} \cdot \Big\{ \int_{\Delta_{k-1}(\gamma_k)} \Big(\prod_{i=1}^{k-1} b_i \mathrm{e}^{-\epsilon_i b_i} \Big) b_k \mathrm{e}^{-(\gamma_k - \epsilon_1 - \ldots - \epsilon_{k-1}) b_k} \mathrm{d}\epsilon_1 \cdots \mathrm{d}\epsilon_{k-1} \Big\} \\ &= \Big\{ \prod_{i=1}^k \frac{x_i^\beta}{b_i} f(x_i) \Big\} \cdot g_{\boldsymbol{b}}(\gamma_k). \end{aligned}$$

This shows that the general gamma density expresses how the $p_{(<)}$-term varies as a function of γ_k. Plugging this into the equation from above, and expanding by $(n-k)!$ and $\phi(\gamma_k)^{n-k}$, yields

$$\begin{aligned} &\frac{n!}{(n-k)!} \prod_{i=1}^k \frac{x_i^\beta}{b_i} f(x_i) \mathrm{d}x_i \cdot \int_0^\infty \phi(\gamma_k)^{n-k} g_{\boldsymbol{b}}(\gamma_k) \\ &\cdot \Big(\int_{\Delta^*_{n-k}(\gamma_k, \infty)} \Big\{ \prod_{i=k+1}^n \frac{x_i^\beta \mathrm{e}^{-\gamma_i x_i^\beta} f(x_i) \mathrm{d}x_i}{-\phi'(\gamma_i)} \Big\} \cdot \Big\{ (n-k)! \prod_{i=k+1}^n \frac{-\phi'(\gamma_i)}{\phi(\gamma_k)} \mathrm{d}\gamma_i \Big\} \Big) \cdot \mathrm{d}\gamma_k \end{aligned} \tag{B.26}$$

Note that (B.26) expresses still the joint pdf of $(X_1, \ldots, X_N)$ given $N = n$. Dividing by $\ell(\boldsymbol{x}|n)\mathrm{d}x_1 \cdots \mathrm{d}x_k$ to get the conditional pdf given $(X_1, \ldots, X_k) = \boldsymbol{x}$, and dropping the integral over γ_k so that this becomes a joint pdf for $(X_{k+1}, \ldots, X_N, \Gamma_k)$, we get

$$\begin{aligned}
&\mathbb{P}[X_{k+1} \in \mathrm{d}x_{k+1}, \ldots, X_n \in \mathrm{d}x_n, \Gamma_k \in \mathrm{d}\gamma_k | (X_1, \ldots, X_k) = \boldsymbol{x}, N = n] \\
&= \left(\frac{\phi(\gamma_k)^{n-k} g_{\boldsymbol{b}}(\gamma_k) \mathrm{d}\gamma_k}{\int_0^\infty \phi(\gamma_k)^{n-k} g_{\boldsymbol{b}}(\gamma_k) \mathrm{d}\gamma_k} \right) \\
&\cdot \int_{\gamma_k}^{\infty} \int_{\gamma_{k+1}}^{\infty} \cdots \int_{\gamma_{n-1}}^{\infty} \left\{ \prod_{i=k+1}^{n} \frac{x_i^\beta \mathrm{e}^{-\gamma_i x_i^\beta} f(x_i) \mathrm{d}x_i}{-\phi'(\gamma_i)} \right\} \cdot \left\{ (n-k)! \prod_{i=k+1}^{n} \frac{-\phi'(\gamma_i)}{\phi(\gamma_k)} \mathrm{d}\gamma_i \right\}.
\end{aligned} \tag{B.27}$$

This shows that $(X_{k+1}, \ldots, X_n)$ depends on the observed data $(X_1, \ldots, X_k) = \boldsymbol{x}$ only via its dependence on the mixing rv Γ_k, which for fixed n has the data-dependent pdf $g^{(0)}(\gamma_k|\boldsymbol{x}, n) \propto \phi(\gamma_k)^{n-k} g_{\boldsymbol{b}}(\gamma_k)$. We also note from comparing (B.24) and (B.27) that the h-density remains unchanged, while the $n-k$ order statistics are now from the truncated cdf $1 - \phi(\gamma)/\phi(\gamma_k)$ defined on $[\gamma_k, \infty)$.

Because the h-density for $X_i|\Gamma_i$ remains unchanged, the Monte Carlo procedure based on inverting decreasing uniform order statistics via the inverse Laplace transform is basically the same as for the "a-priori case". For the procedure based on exponential order statistics, it is now required that the $(\tilde{X}_{k+1}, \ldots, \tilde{X}_n)$ are drawn from the remaining data pdf $\rho(\tilde{x}|\gamma_k)$ defined in equation (4.15), before proceeding as in the "a-priori case" to obtain $X_{k+1} = \tilde{X}_{i(k+1)}, \ldots, X_n = \tilde{X}_{i(n)}$.

An accept / reject procedure for drawing from the h-density

For the case of $w(x) = x^\beta$ (with $\beta \neq 0$), it is possible to make a rearrangement which relates the h-density to the pdf of the Weibull distribution. Let us denote the sign of β by $d = \mathrm{sign}\beta$, so that $w(x) = x^{d|\beta|}$. Then, the h-density is proportional to

$$\begin{aligned}
h(x|\gamma) &\propto x^{d|\beta|} \mathrm{e}^{-\gamma x^{d|\beta|}} f(x) \\
&\propto \gamma|\beta| x^{d(|\beta|-1)} \mathrm{e}^{-\gamma x^{d|\beta|}} \cdot x^d f(x), \ \forall x \in [\underline{x}, \overline{x}].
\end{aligned} \tag{B.28}$$

Letting $h^{(\mathrm{Wei})}_{(|\beta|,\gamma)}(s) = \gamma|\beta| s^{|\beta|-1} \mathrm{e}^{-\gamma s^{|\beta|}}$ denote the pdf of the Weibull distribution with shape parameter $|\beta|$ and scale parameter γ, (B.28) shows that $h(x|\gamma) \propto h^{(\mathrm{Wei})}_{(|\beta|,\gamma)}(x^d) \cdot x^d f(x)$. The parameterization adopted here with the scale parameter γ is the one often encountered in econometrics (e.g. Wooldridge (2002), p. 689).[101]

Draws from the h-density for $X_i|\Gamma_i$ are easily obtained by an accept / reject procedure (e.g. Robert and Casella (2010)). Consider first the case of $\beta > 0$. Let X denote repeated drawings from the truncated Wei$(|\beta|, \gamma_i)$-distribution on $[\underline{x}, \overline{x}]$, and U from the U$(0,1)$-distribution. Then, since the function $h^{(\mathrm{Wei})}_{(|\beta|,\gamma_i)}(x) \cdot \max_{r \in [\underline{x}, \overline{x}]} \{rf(r)\}$ envelopes $h^{(\mathrm{Wei})}_{(|\beta|,\gamma_i)}(x) \cdot xf(x)$, the rule for accepting becomes as follows: accept X as a draw for X_i for the first pair (X, U) which fulfills the inequality $U \cdot \max_{r \in [\underline{x}, \overline{x}]} \{rf(r)\} < Xf(X)$.[102]

[101] An alternative parameterization encountered in statistics is based on substituting $\zeta = \gamma^{-1/|\beta|}$ as the scale parameter, which is then directly proportional to the mean of the Weibull distribution. This substitution might be impractical if $|\beta|$ is very small.

[102] Note that for $\beta > 0$ and for a truncated log-uniform field-size distribution, which has density $f(x) \propto 1/x$ on $[\underline{x}, \overline{x}]$, the result is that $X_i|\Gamma_i$ has the truncated Wei(β, γ_i)-distribution on $[\underline{x}, \overline{x}]$.

In the case of $\beta < 0$ we have $d = -1$, so that we get $h^{(\text{Wei})}_{(|\beta|,\gamma_i)}(x^{-1}) \cdot x^{-1} f(x) = h^{(\text{Wei})}_{(|\beta|,\gamma_i)}(x^{-1}) x^{-2} \cdot xf(x)$. In fact, integrating gives $\int_{\underline{x}}^{x} h^{(\text{Wei})}_{(|\beta|,\gamma_i)}(r^{-1}) r^{-2} \mathrm{d}r = \int_{\overline{x}^{-1}}^{\underline{x}^{-1}} h^{(\text{Wei})}_{(|\beta|,\gamma_i)}(s) \mathrm{d}s$, so that we can make the repeated drawings X^{-1} from the truncated $\text{Wei}(|\beta|, \gamma_i)$-distribution on support $[\overline{x}^{-1}, \underline{x}^{-1}]$. We then accept the inverse of the draw, X, as a draw for X_i for the first pair (X^{-1}, U) which fulfills $U \cdot \max_{r \in [\underline{x}, \overline{x}]} \{rf(r)\} < Xf(X)$.

C Appendix to Chapter 4

Figure C1: Actual oil (black) and gas (red) production for the three regions

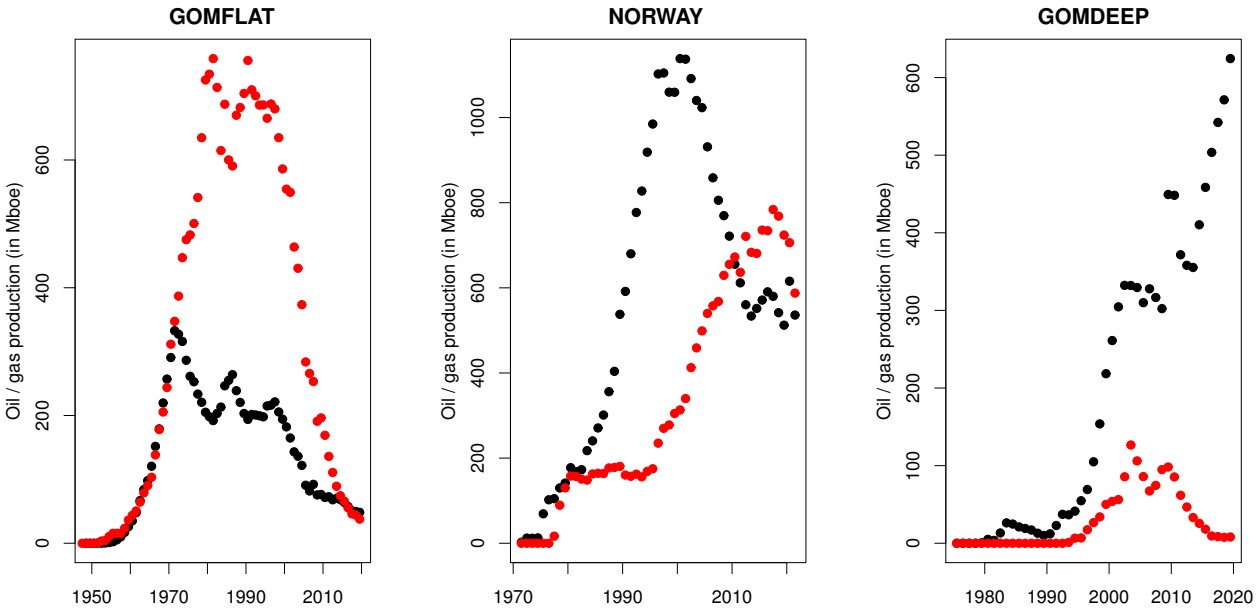
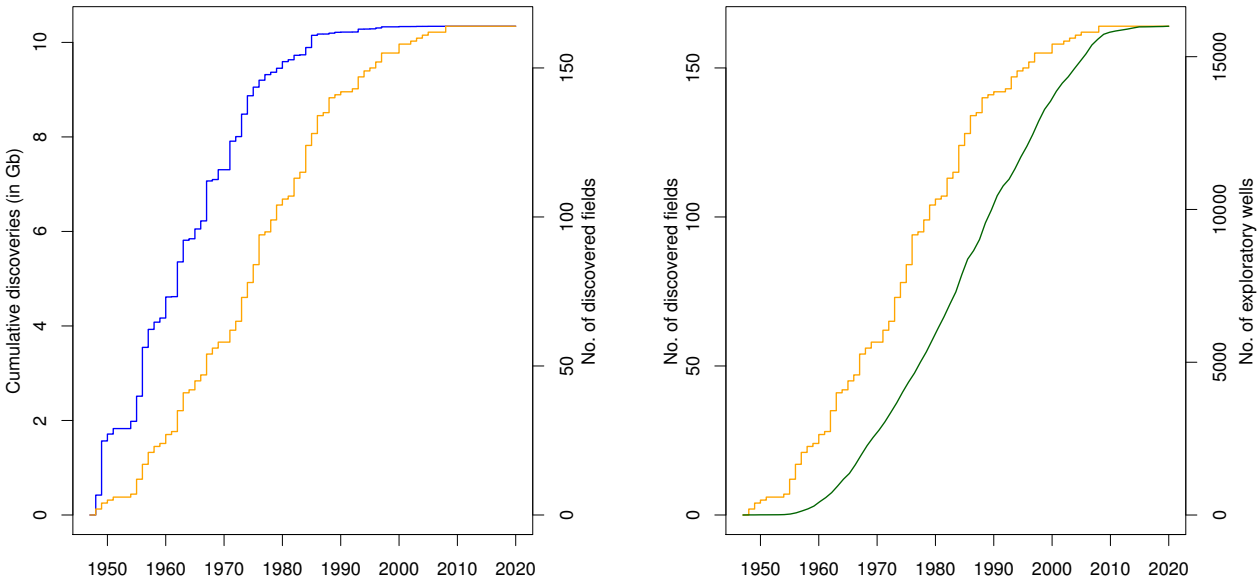
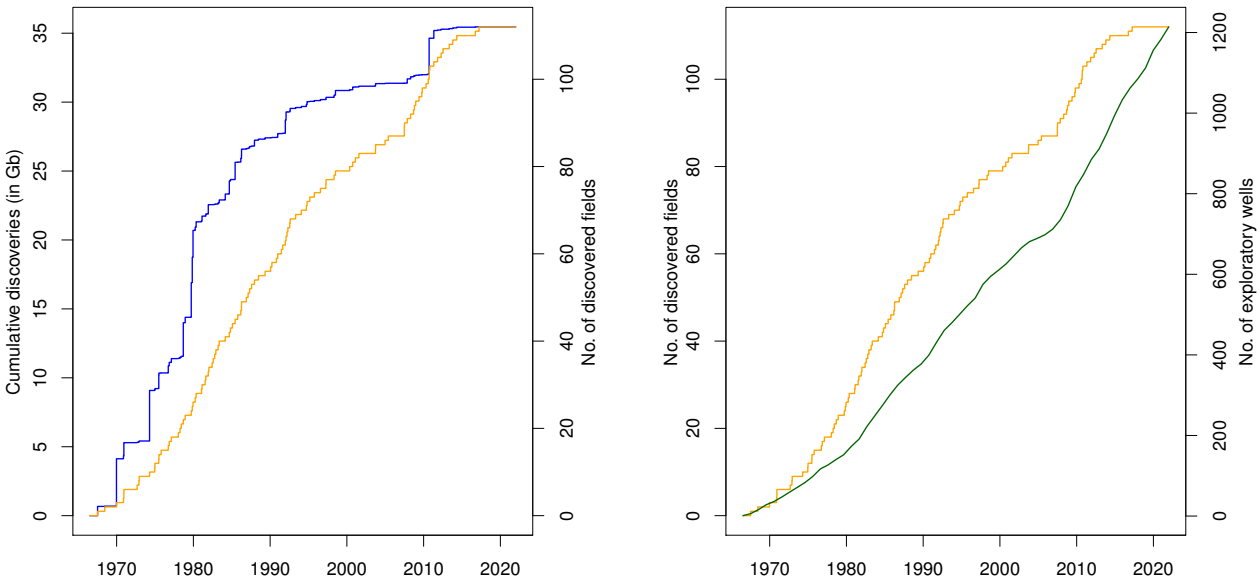


Figure C2: Discovery and exploration history - GOM Flat



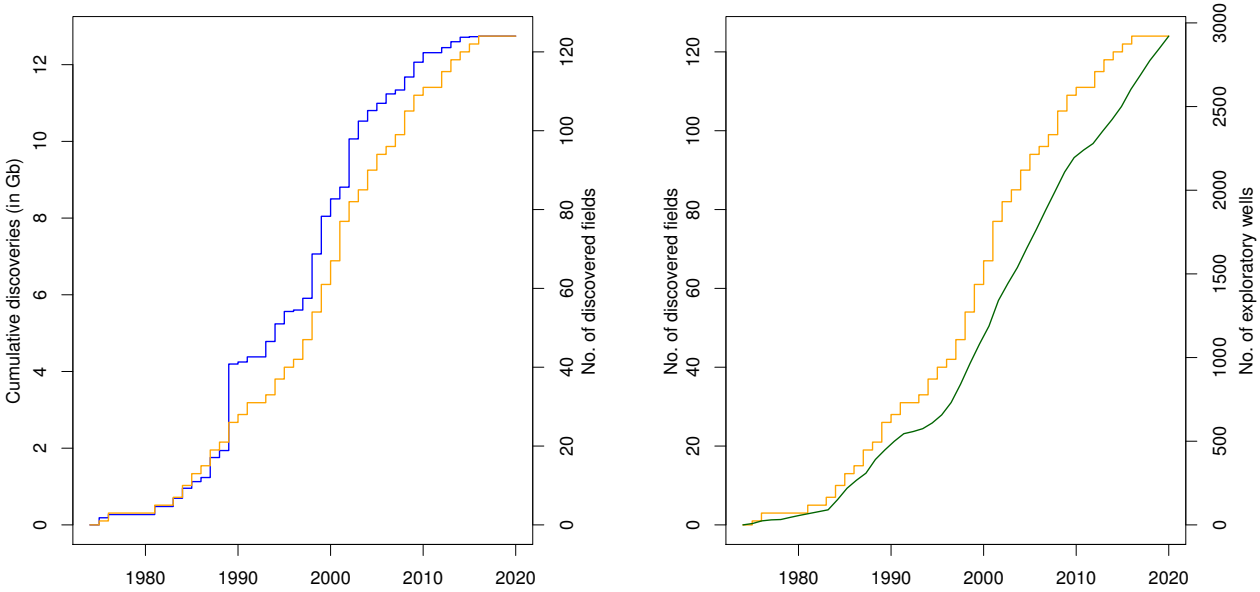
Note: depicted are the cumulative amounts of discovered oil (blue), the cumulative number of discovered oil fields (orange), and the cumulative number of drilled exploratory wells (green)

Figure C3: Discovery and exploration history - Norway



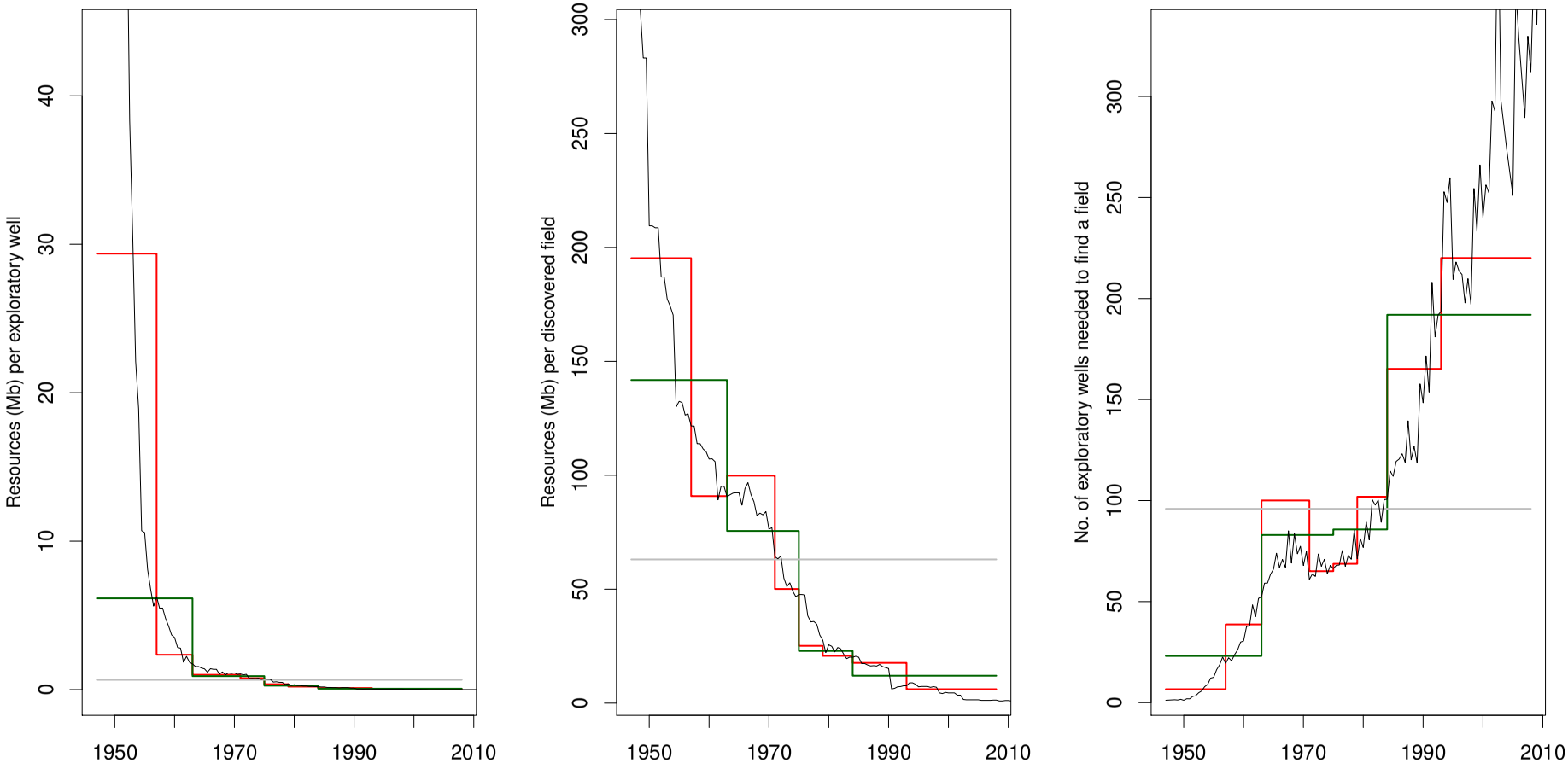
Note: depicted are the cumulative amounts of discovered oil (blue), the cumulative number of discovered oil fields (orange), and the cumulative number of drilled exploratory wells (green)

Figure C4: Discovery and exploration history - GOM Deep



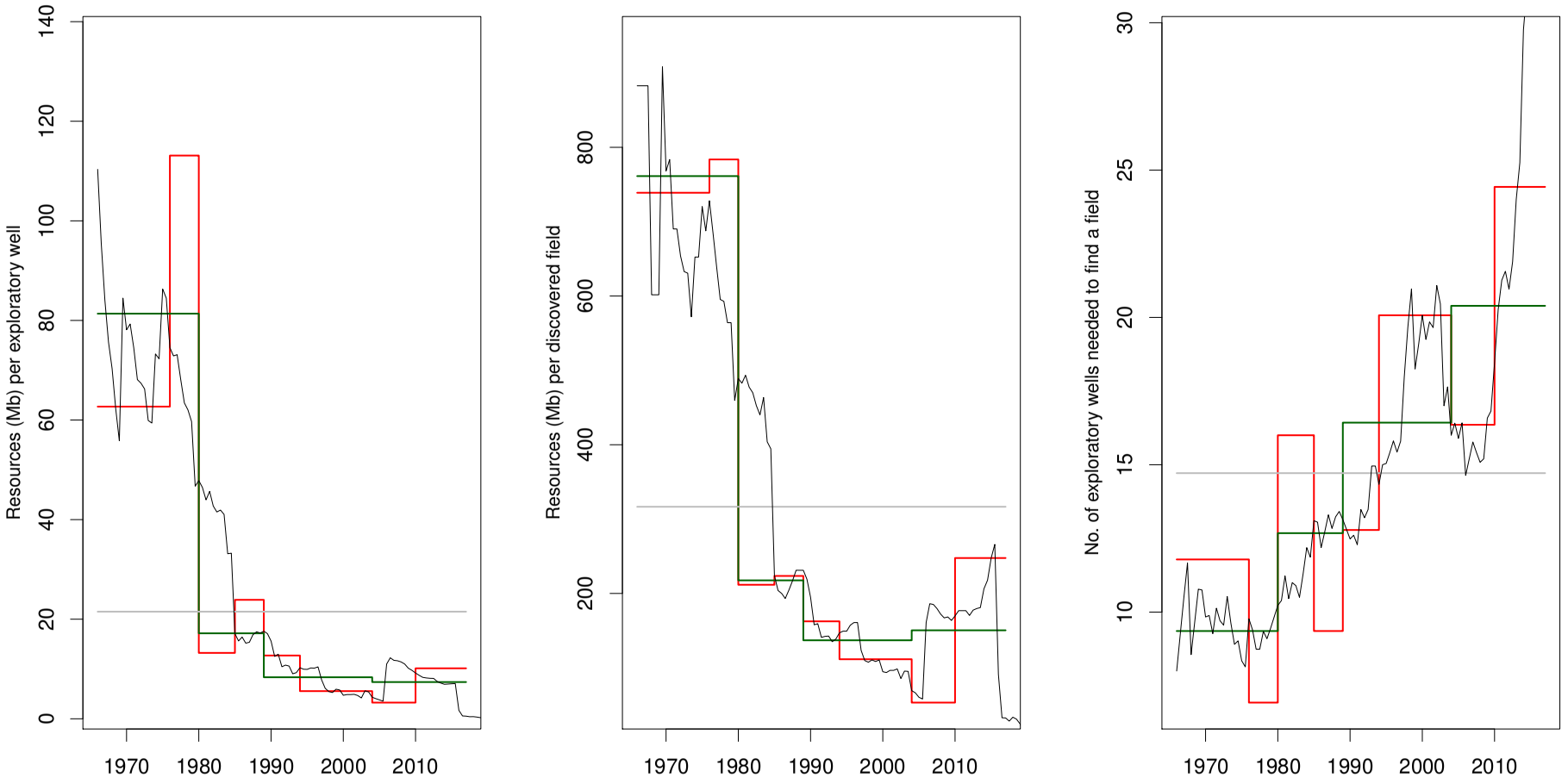
Note: depicted are the cumulative amounts of discovered oil (blue), the cumulative number of discovered oil fields (orange), and the cumulative number of drilled exploratory wells (green)

Figure C5: Illustration of the depletion effect - GOM Flat



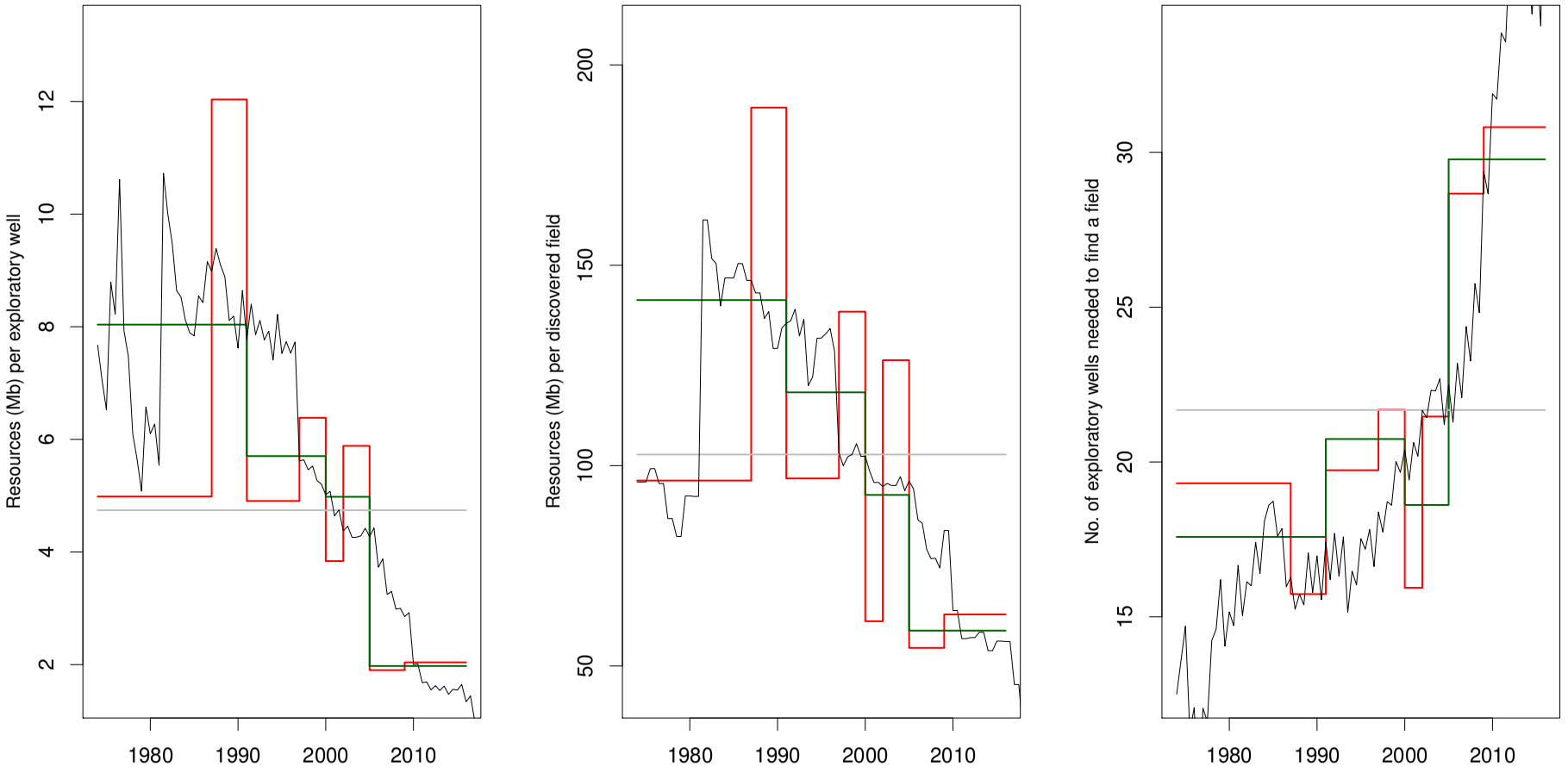
Note: the gray line depicts the average over all discovered fields, while for the green and the red lines the sample is divided into 4 and 8 equally-sized groups, respectively. The black line shows a moving average computed using a time window with length of 10 years.

Figure C6: Illustration of the depletion effect - Norway



Note: the gray line depicts the average over all discovered fields, while for the green and the red lines the sample is divided into 4 and 8 equally-sized groups, respectively. The black line shows a moving average computed using a time window with length of 10 years.

Figure C7: Illustration of the depletion effect - GOM Deep



Note: the gray line depicts the average over all discovered fields, while for the green and the red lines the sample is divided into 4 and 8 equally-sized groups, respectively. The black line shows a moving average computed using a time window with length of 10 years.

Table C1: Estimation results for $\hat{a}_0$ and $\hat{b}_0$

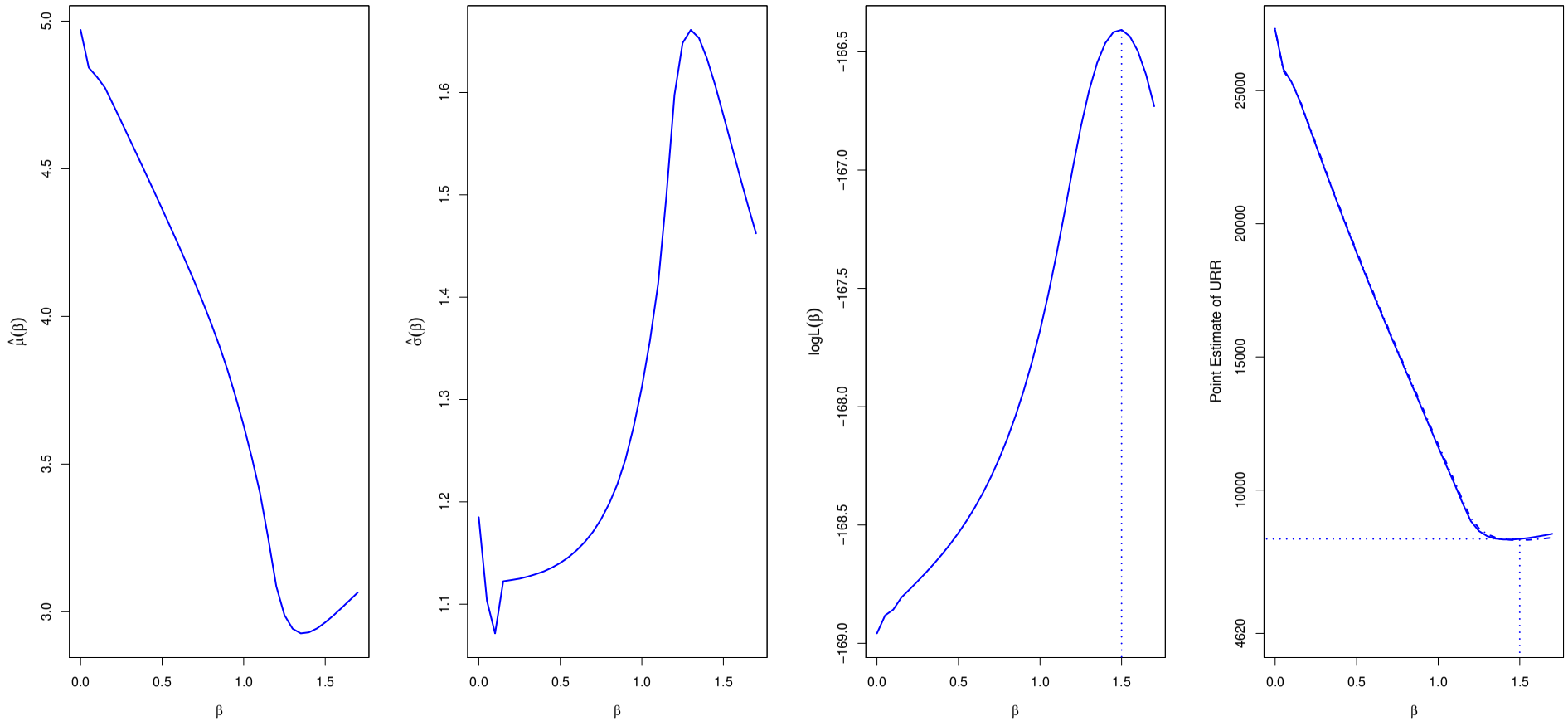
	GOM Flat, $\underline{t} = 1962$		GOM Flat, $\underline{t} = 1967$		Norway, $\underline{t} = 1984$		Norway, $\underline{t} = 1989$	
$\overline{x}$	600	600	600	600	4400	4400	4400	4400
ν	164	164	164	164	150	150	150	150
	ML-EST	FIT-URR	ML-EST	FIT-URR	ML-EST	FIT-URR	ML-EST	FIT-URR
e_{k+1}	27308	8686	1772	812	628	311	526	221
$\hat{a}_0$	1.00	1.00	1.00	1.00	7.14	5.59	1.70	1.00
	(NA)[a]	(NA)	(NA)	(NA)	(1.82)	(3.72)	(1.91)	(NA)
$\frac{\hat{b}_0}{\hat{a}_0 e_{k+1}}$	19.69	17.20	45.22	32.50	0.27	0.62	7.71	11.93
	(6.29)	(5.82)	(4.06)	(3.26)	(0.36)	(1.14)	(10.20)	(1.09)
R^2	0.21	0.10	0.52	0.31	0.05	0.05	0.51	0.42

Note: HAC (Newey/West) standard errors are in parenthesis.
[a]Whenever $\hat{a}_0 < 1$ is obtained we set $\hat{a}_0 = 1$, hence no standard errors are available.

	Norway, $\underline{t} = 2022$		GOM Deep, $\underline{t} = 2000$		GOM Deep, $\underline{t} = 2005$		GOM Deep, $\underline{t} = 2020$	
$\overline{x}$	4400	4400	700	700	700	700	700	700
ν	150	150	300	300	300	300	300	300
	ML-EST	FIT-URR	ML-EST	FIT-URR	ML-EST	FIT-URR	ML-EST	FIT-URR
e_{k+1}	95	38	64	102	644	58	561	25
$\hat{a}_0$	3.97	1.00	1.00	1.00	3.88	1.77	1.96	1.00
	(3.05)	(NA)[a]	(NA)	(NA)	(4.97)	(5.96)	(2.08)	(NA)
$\frac{\hat{b}_0}{\hat{a}_0 e_{k+1}}$	7.26	29.27	12.36	12.15	2.88	7.54	10.81	22.11
	(8.08)	(3.80)	(1.12)	(1.12)	(5.22)	(29.34)	(13.06)	(1.00)
R^2	0.33	0.30	0.14	0.13	0.17	0.18	0.58	0.57

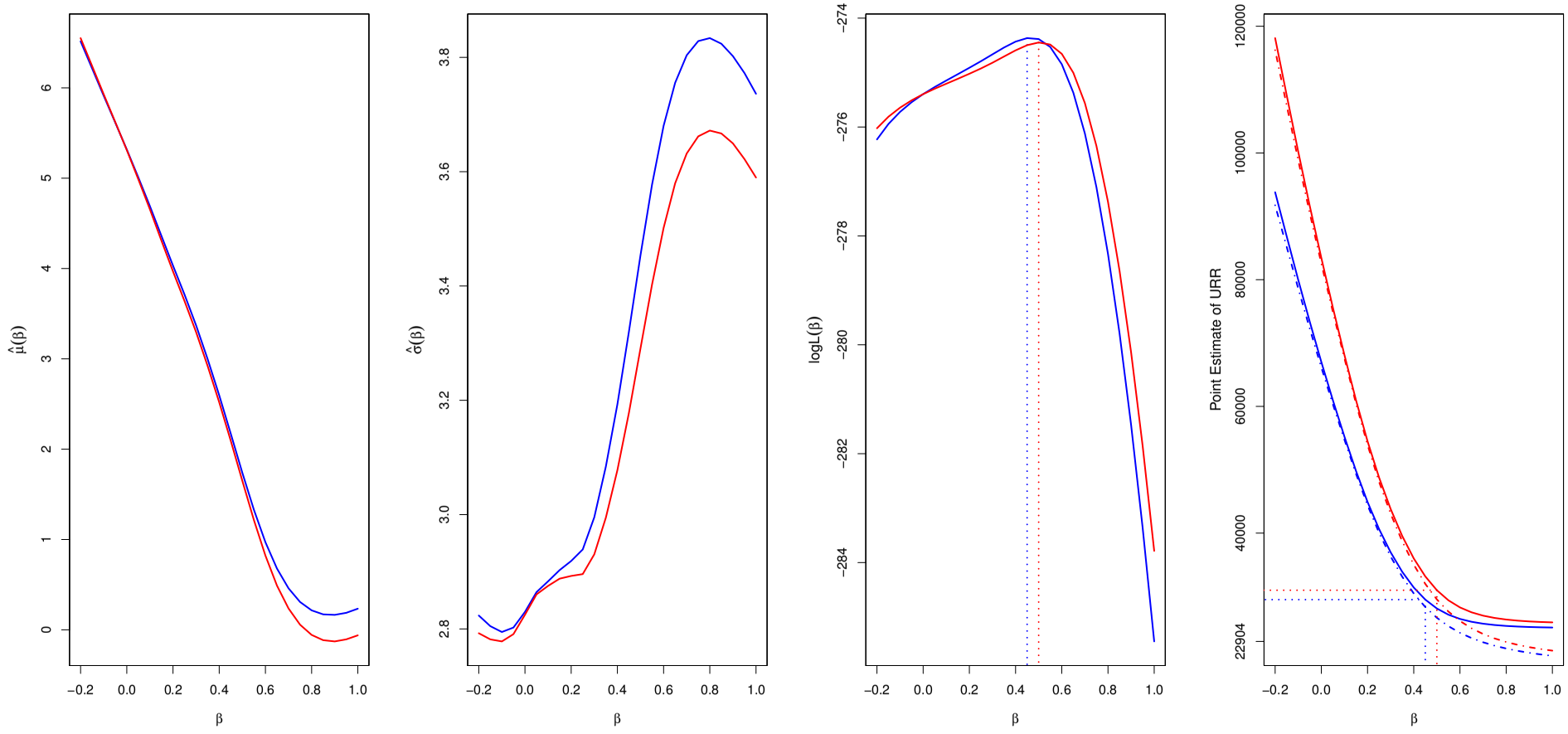
Note: HAC (Newey/West) standard errors are in parenthesis.
[a]Whenever $\hat{a}_0 < 1$ is obtained we set $\hat{a}_0 = 1$, hence no standard errors are available.

Figure C8: Estimation results for GOM Flat ($\underline{t}$ =1962)



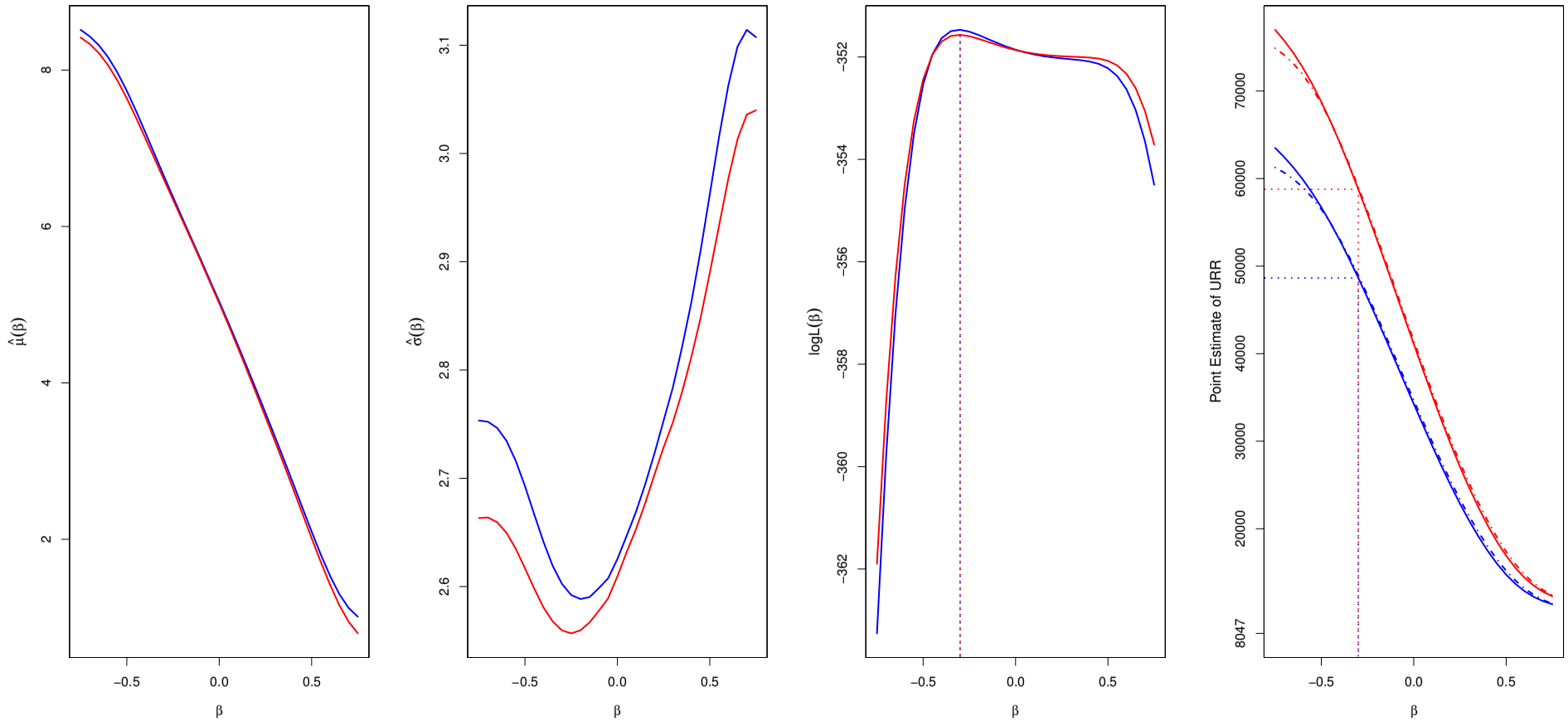
Note: the curves depict estimation results for $\nu = 164$. In the last panel the dot-dashed lines show the prior mean, while the solid lines show the posterior mean including the observed field sizes.

Figure C9: Estimation results for Norway ($\underline{t}$ =1984)



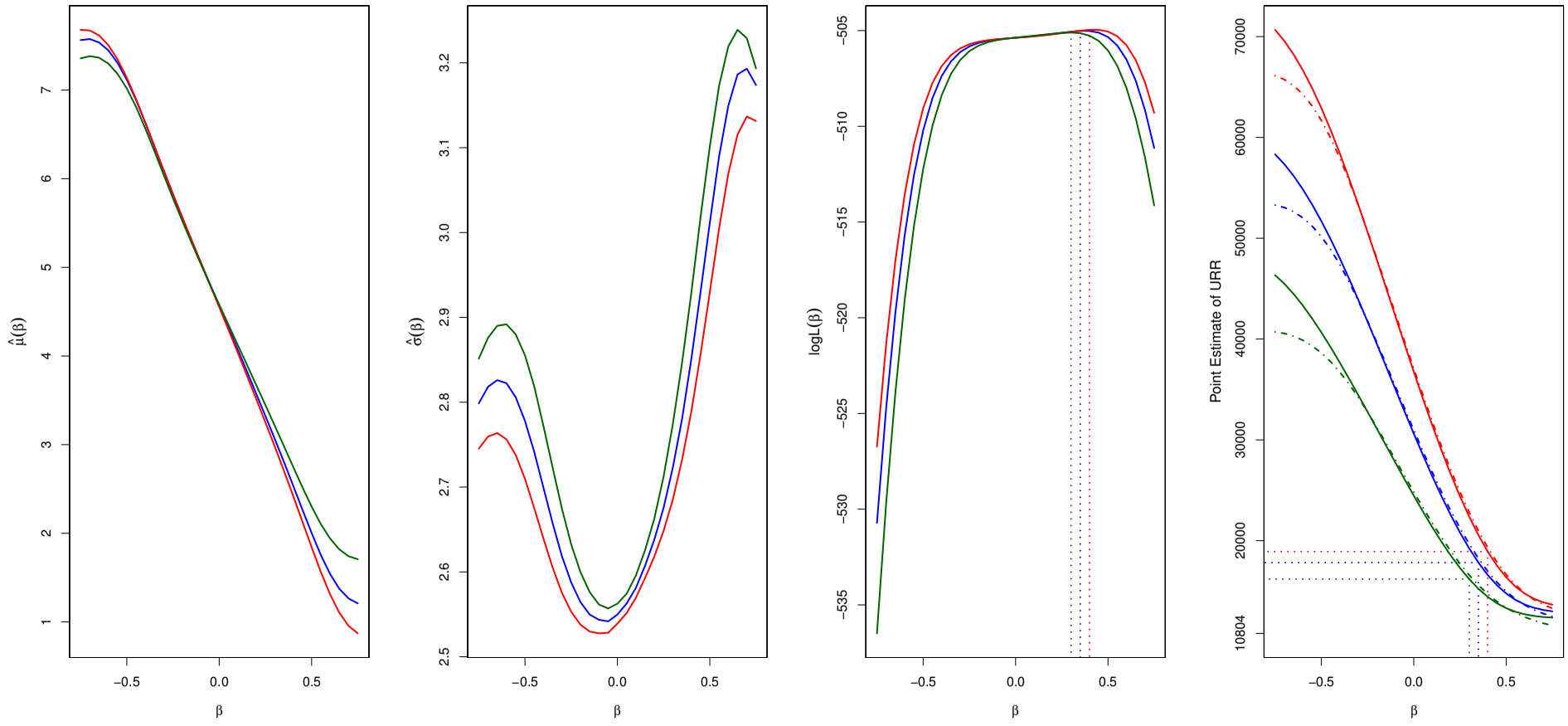
Note: the curves depict estimation results for $\nu = 150$ (in red) and $\nu = 120$ (in blue). In the last panel the dot-dashed lines show the prior mean, while the solid lines show the posterior mean including the observed field sizes.

Figure C10: Estimation results for GOM Deep ($\underline{t}$ =2000)



Note: the curves depict estimation results for $\nu = 300$ (in red) and $\nu = 250$ (in blue). In the last panel the dot-dashed lines show the prior mean, while the solid lines show the posterior mean including the observed field sizes.

Figure C11: Estimation results for GOM Deep ($\underline{t}$ =2005)



Note: the curves depict estimation results for $\nu = 300$ (in red), $\nu = 250$ (in blue), and $\nu = 200$ (in green). In the last panel the dot-dashed lines show the prior mean, while the solid lines show the posterior mean including the observed field sizes.

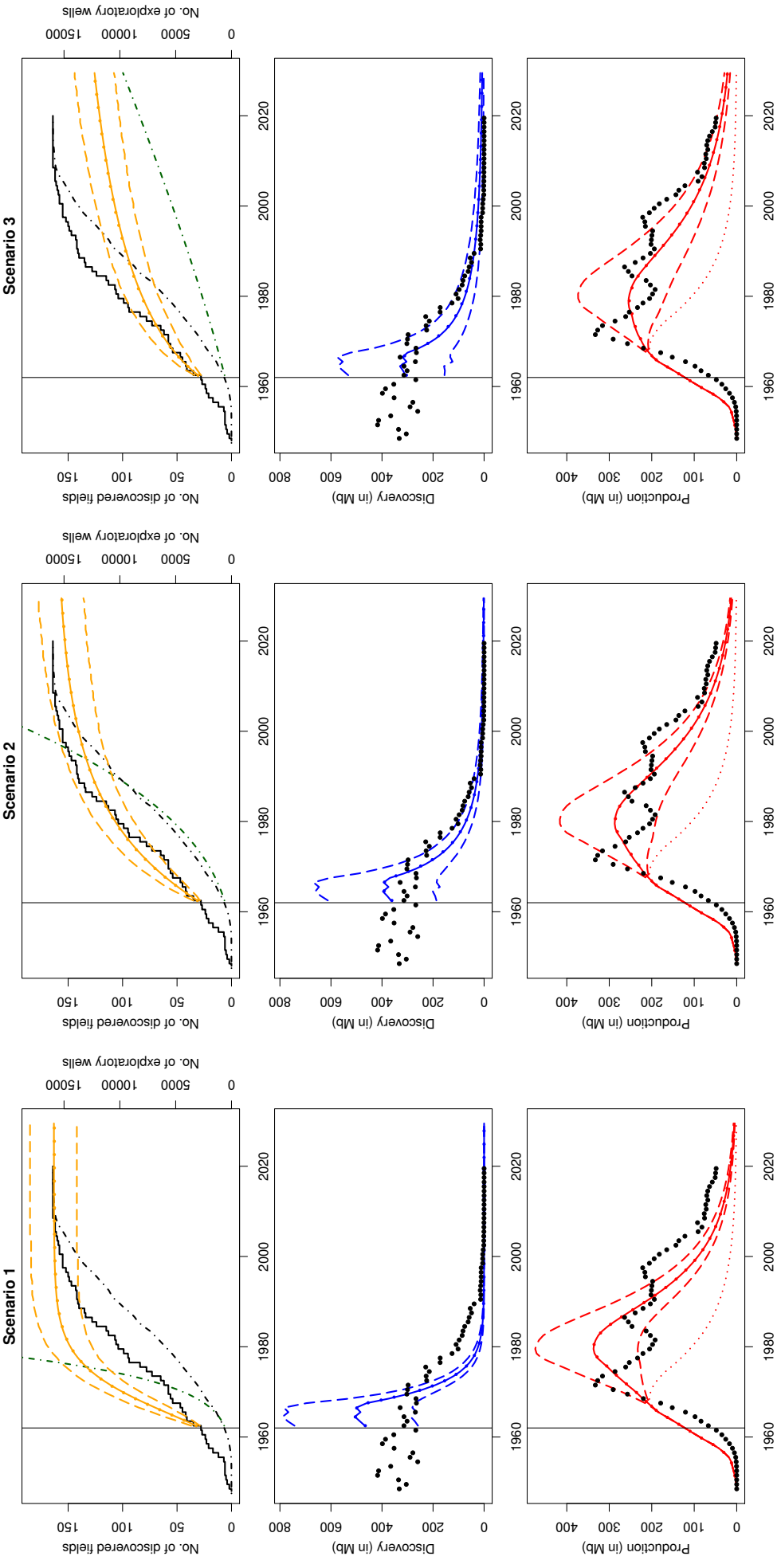


Figure C12: Simulation results for GOM Flat ($\underline{t} = 1962$, FIT-URR, assumption (4b))

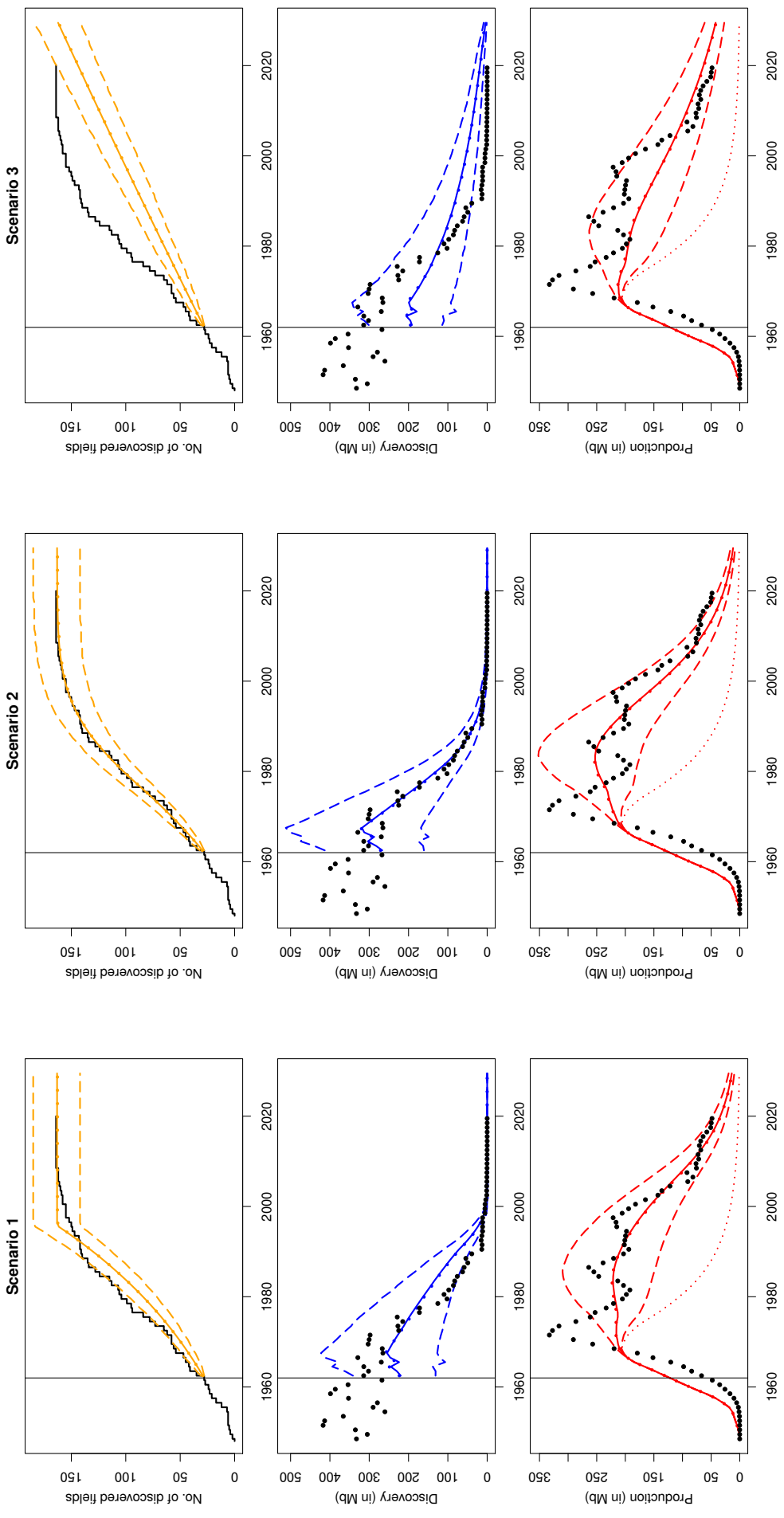
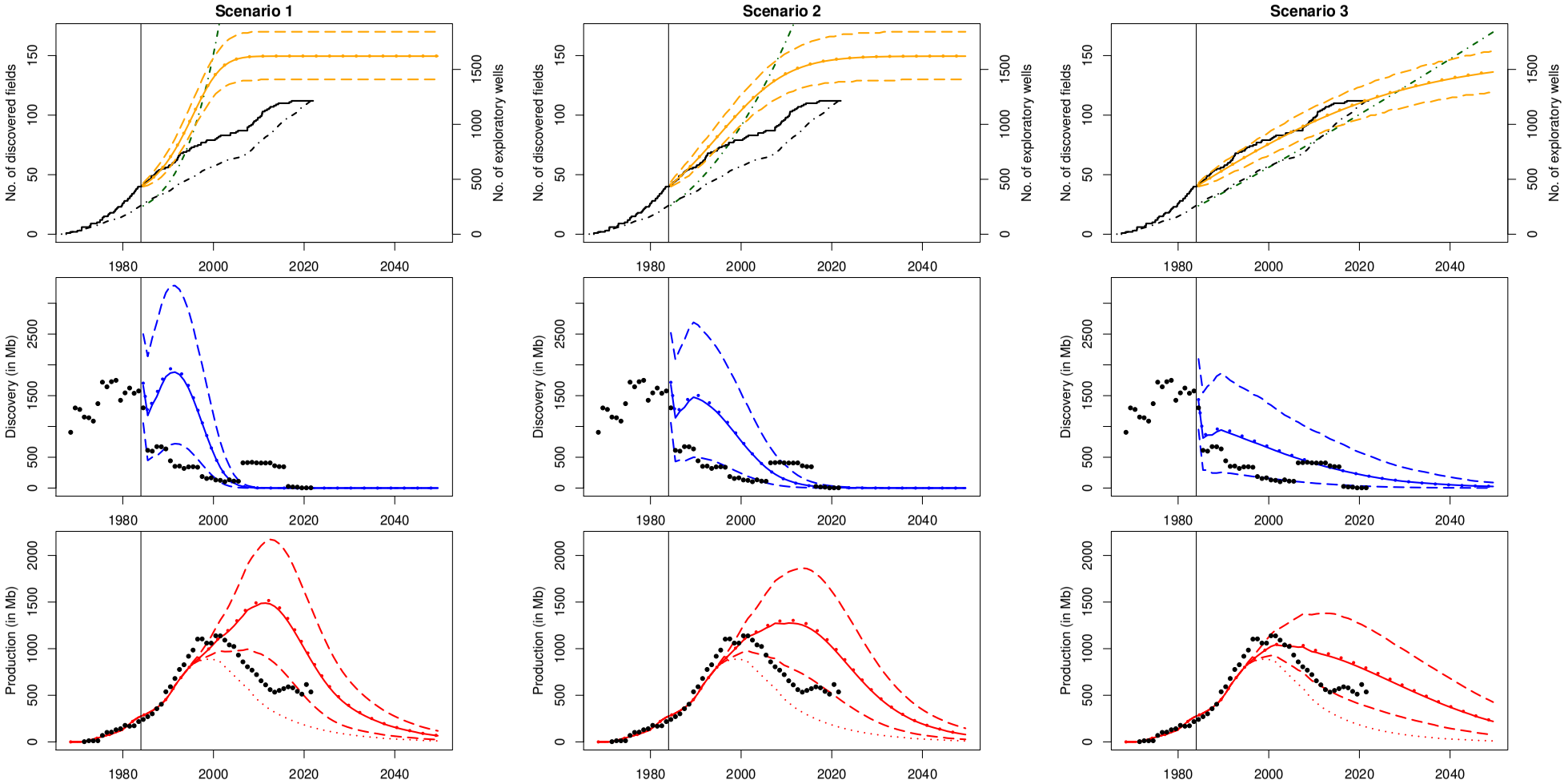


Figure C13: Simulation results for GOM Flat ($\underline{t} = 1962$, FIT-URR, assumption (4a))

Figure C14: Simulation results for Norway ($\underline{t} = 1984$, FIT-URR, assumption (4b))



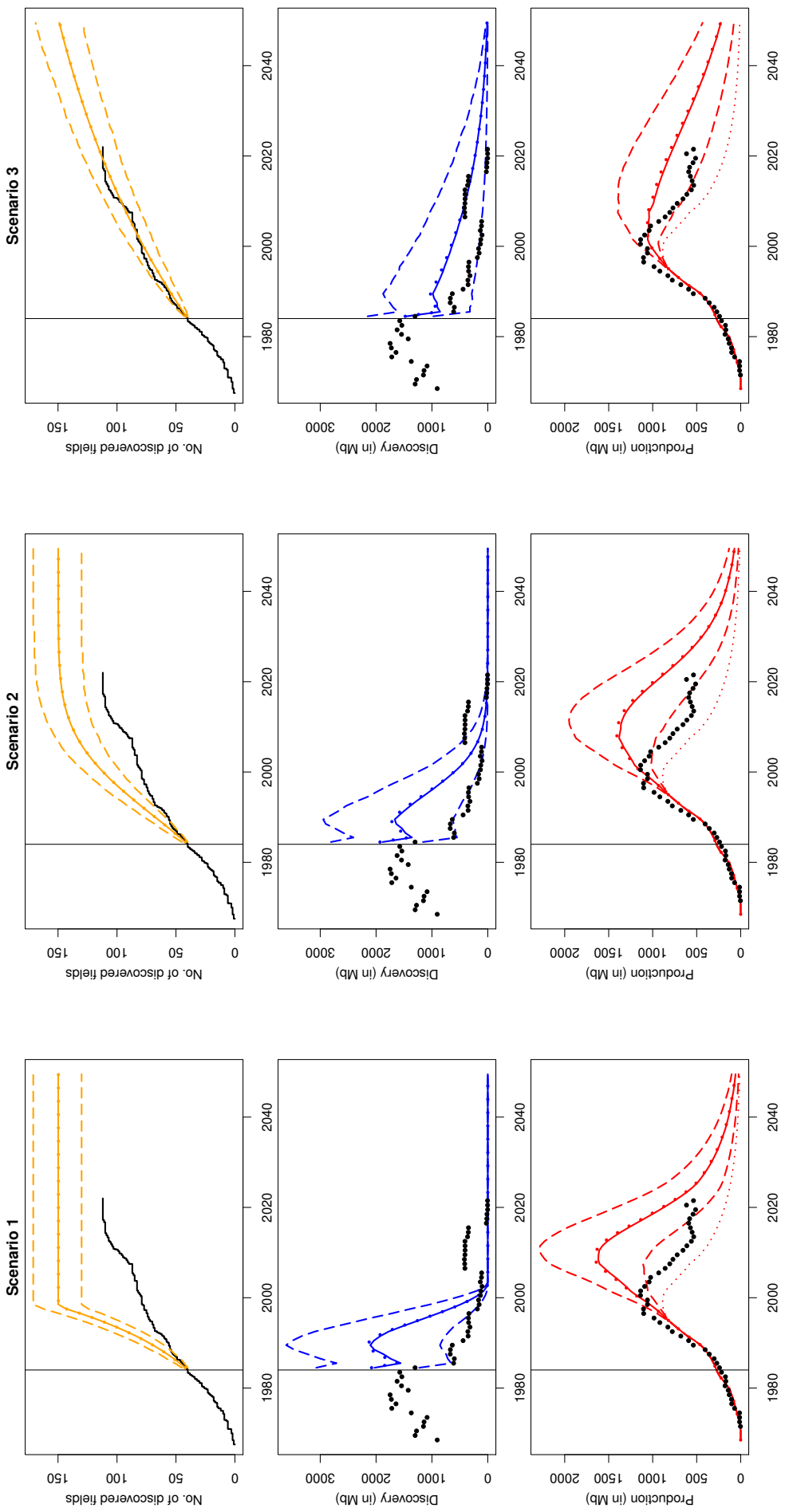


Figure C15: Simulation results for Norway ($\underline{t} = 1984$, FIT-URR, assumption (4a))

Figure C16: Simulation results for GOM Deep ($\underline{t} = 2005$, ML-EST, assumption (4b))

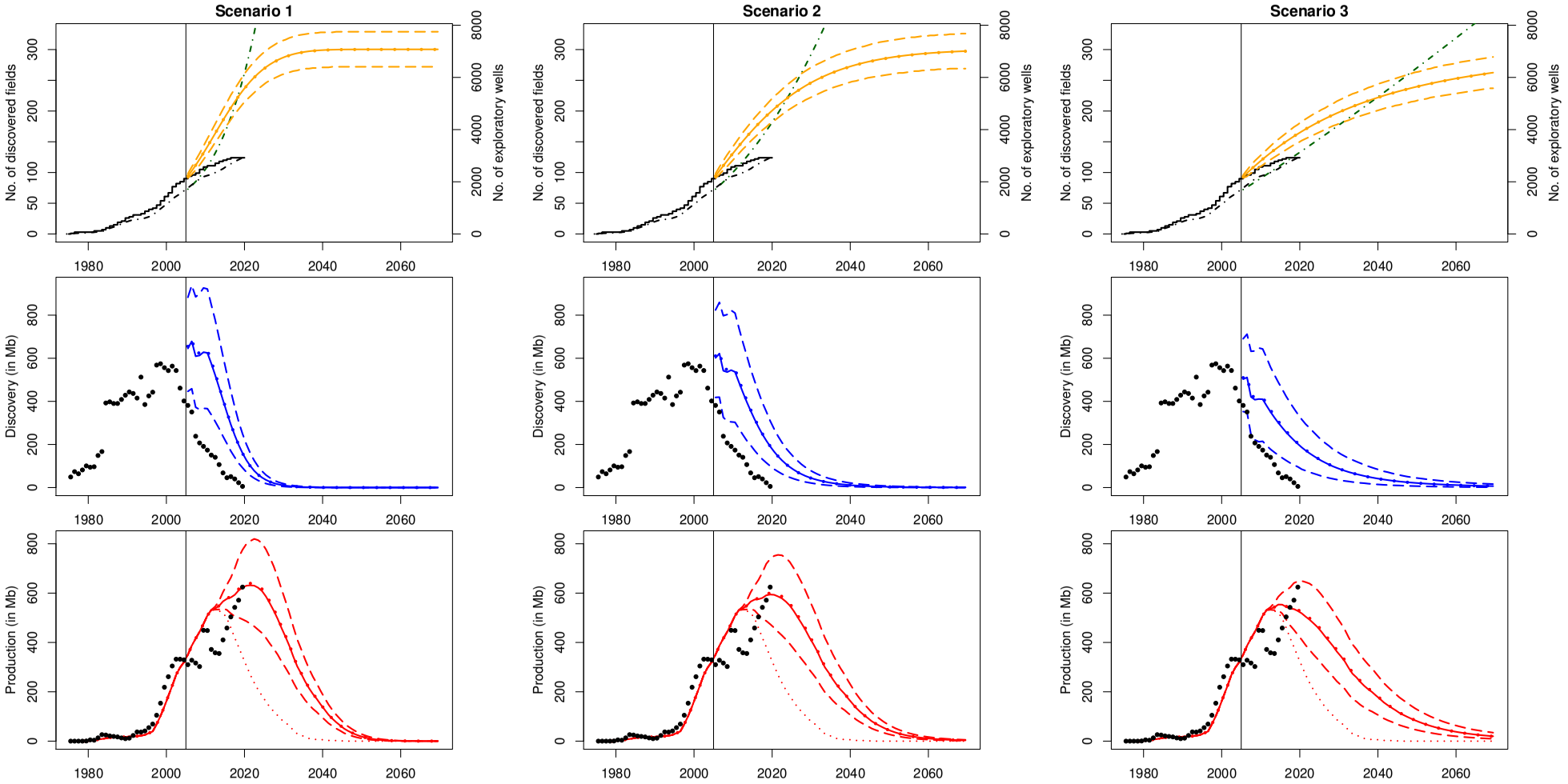
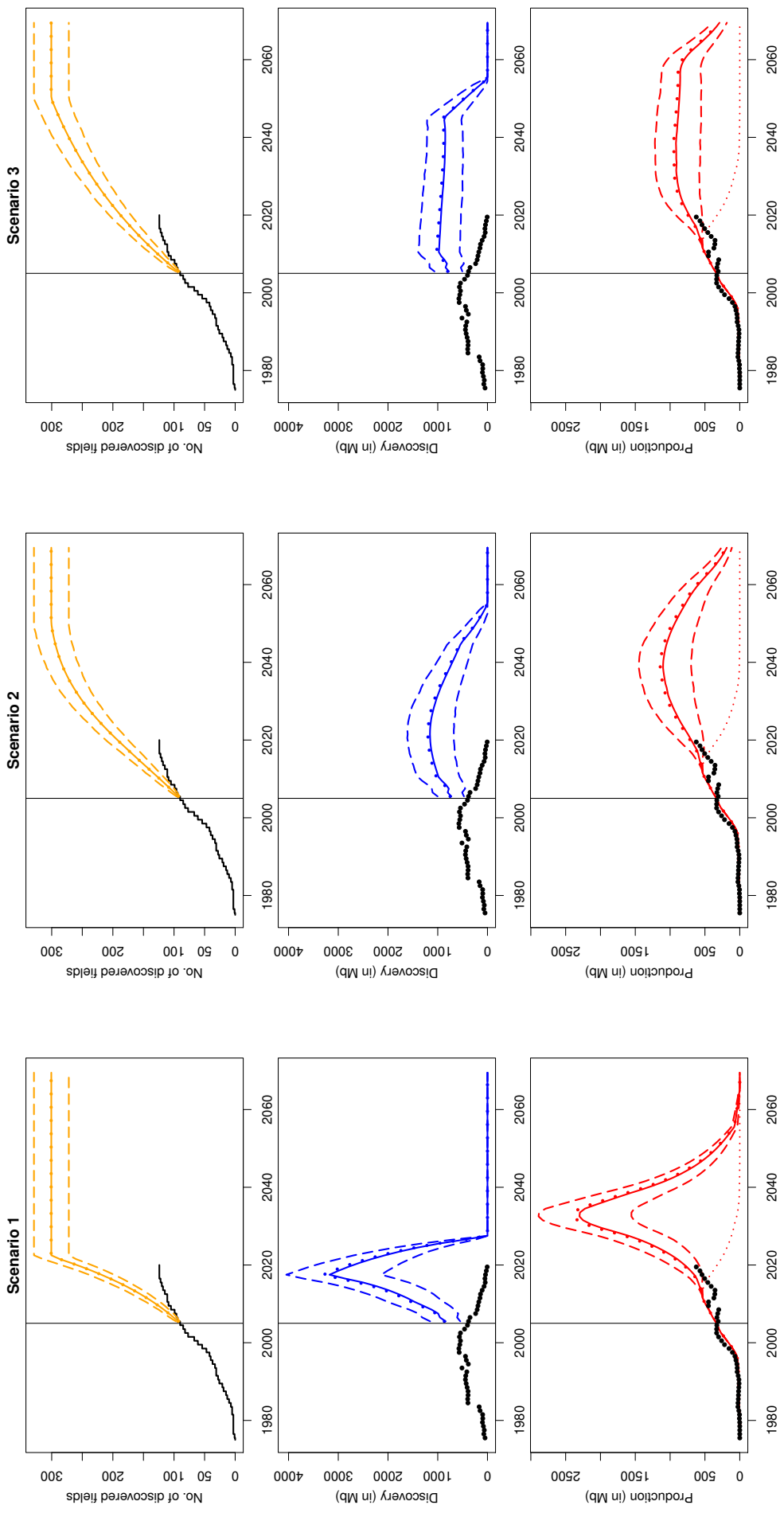


Figure C17: Simulation results for GOM Deep ($\underline{t} = 2005$, FIT-URR, assumption (4a))



D Appendix to Chapter 5

Table D1: Results of unit-root tests

Series	Levels				First differences			
	Trend?	ADF	ERS	KPSS	Trend?	ADF	ERS	KPSS
ELP	No	-3.37**	-3.07***	4.551***	No	-12.92***	-9.30***	0.0229
COP	No	-1.42	-1.03	4.935***	No	-27.23***	-16.29***	0.1065
NGP	No	-2.24	-2.26**	2.192***	No	-28.82***	-14.03***	0.0625
RES	No	-3.38**	-1.18	5.845***	No	-13.31***	-0.22	0.0164
TOT	No	-3.76***	-2.79***	1.505***	No	-10.94***	-2.57***	0.0086
VRE	No	-20.64***	-13.25***	5.307***	No	-23.55***	-2.73***	0.0139

Note: *p<0.1,**p<0.05,***p<0.01. ADF=Augmented Dickey-Fuller test. No. of lags selected via BIC. ERS=Elliott/Rothenberg/Stock test. No. of lags is the same number as used for the ADF test. KPSS=Kwiatkowski/Phillips/Schmidt/Shin test. No. of lags selected as 'long' in R-method 'ur.kpss'.

Table D2: German VRE and nuclear capacity, mean and variability of daily generation

		2015	2016	2017	2018	2019	2020	2021	2022	2023
Photovoltaic	Installed generating capacity[a] [GW]	38.0	39.8	41.8	44.1	46.8	50.8	55.5	60.4	68.4
	Daily generation, mean [GWh]	116	116	116	132	133	146	147	172	176
	Daily generation, capacity factor [%]	13	12	12	12	12	12	11	12	11
	Daily generation, coef. of var. (CV) [%]	55	55	58	59	60	58	60	57	60
	Daily generation, robust CV[b] [%]	48	47	49	54	52	51	52	53	54
Wind onshore	Installed generating capacity[a] [GW]	39.4	44.1	49.3	52.2	53.0	53.8	54.9	56.4	58.8
	Daily generation, mean [GWh]	186	178	233	245	274	282	245	276	323
	Daily generation, capacity factor [%]	20	17	20	20	22	22	19	20	23
	Daily generation, coef. of var. (CV) [%]	79	77	74	72	71	73	76	77	72
	Daily generation, robust CV[b] [%]	53	48	56	50	48	52	48	55	53
Wind offshore	Installed generating capacity[a] [GW]	2.1	3.7	4.6	5.7	6.9	7.6	7.8	8.0	8.3
	Daily generation, mean [GWh]	38	44	54	55	69	77	71	75	68
	Daily generation, capacity factor [%]	73	49	49	40	41	42	38	39	34
	Daily generation, coef. of var. (CV) [%]	155	99	62	59	51	57	62	60	58
	Daily generation, robust CV[b] [%]	61	51	43	46	44	51	55	47	47
Nuclear	Installed generating capacity[a] [GW]	11.4	10.8	10.2	9.5	8.8	8.1	6.1	4.1	2.0
	Daily generation, mean [GWh]	231	219	198	197	195	166	179	90	19
	Daily generation, capacity factor [%]	84	85	81	86	92	85	123	92	38
	Daily generation, coef. of var. (CV) [%]	13	16	17	12	15	14	8	13	158
	Daily generation, robust CV[b] [%]	8	14	18	9	10	10	7	4	100

[a]Reported are the means between two consecutive years as the original data refer to installed capacity on 1st of January.
[b]Defined as the interquartile range divided by the sum of the quartiles.
Source: German Federal Network Agency (Bundesnetzagentur), https://www.smard.de/

Figure D1: Univariate wavelet results for time series in levels

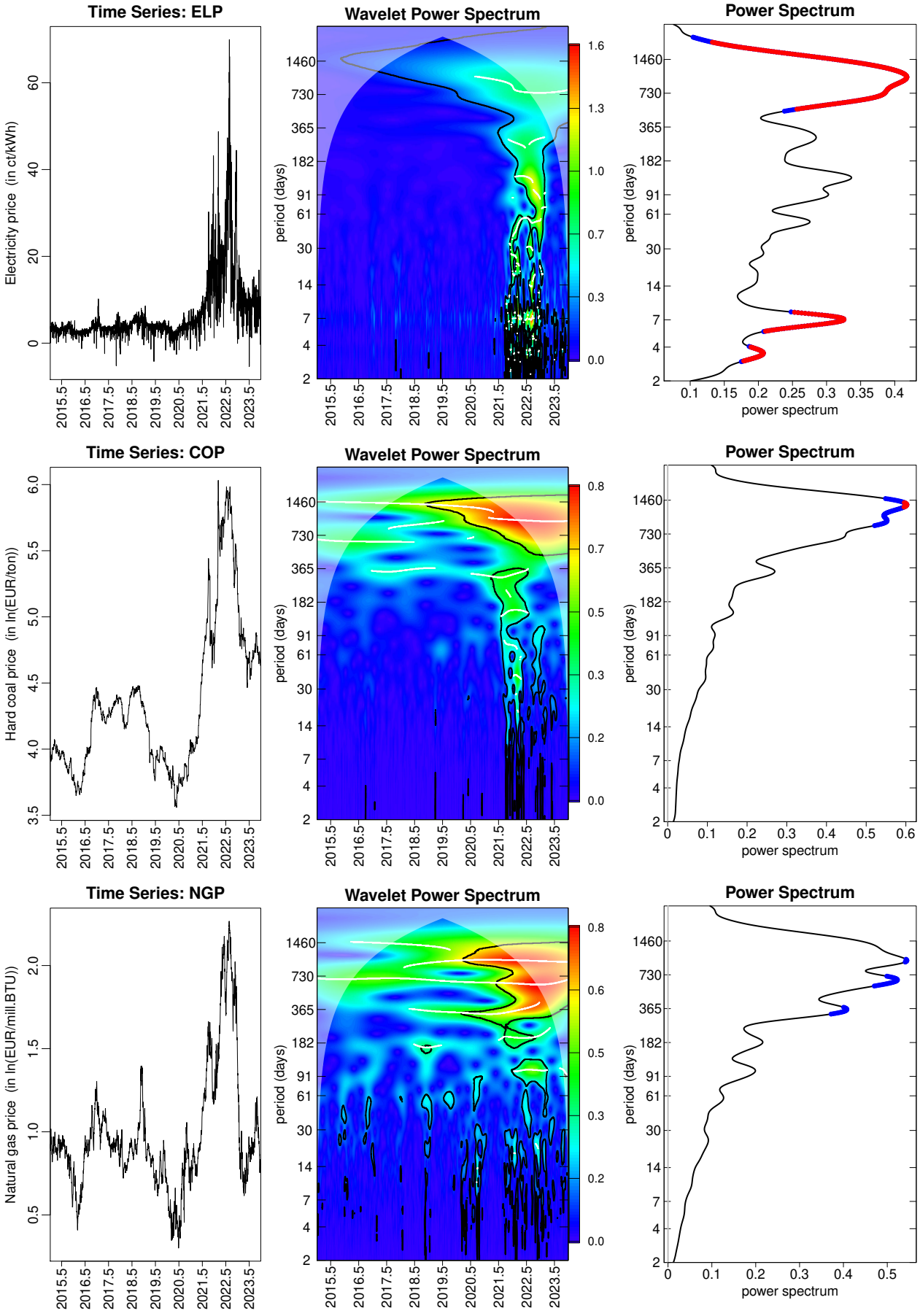
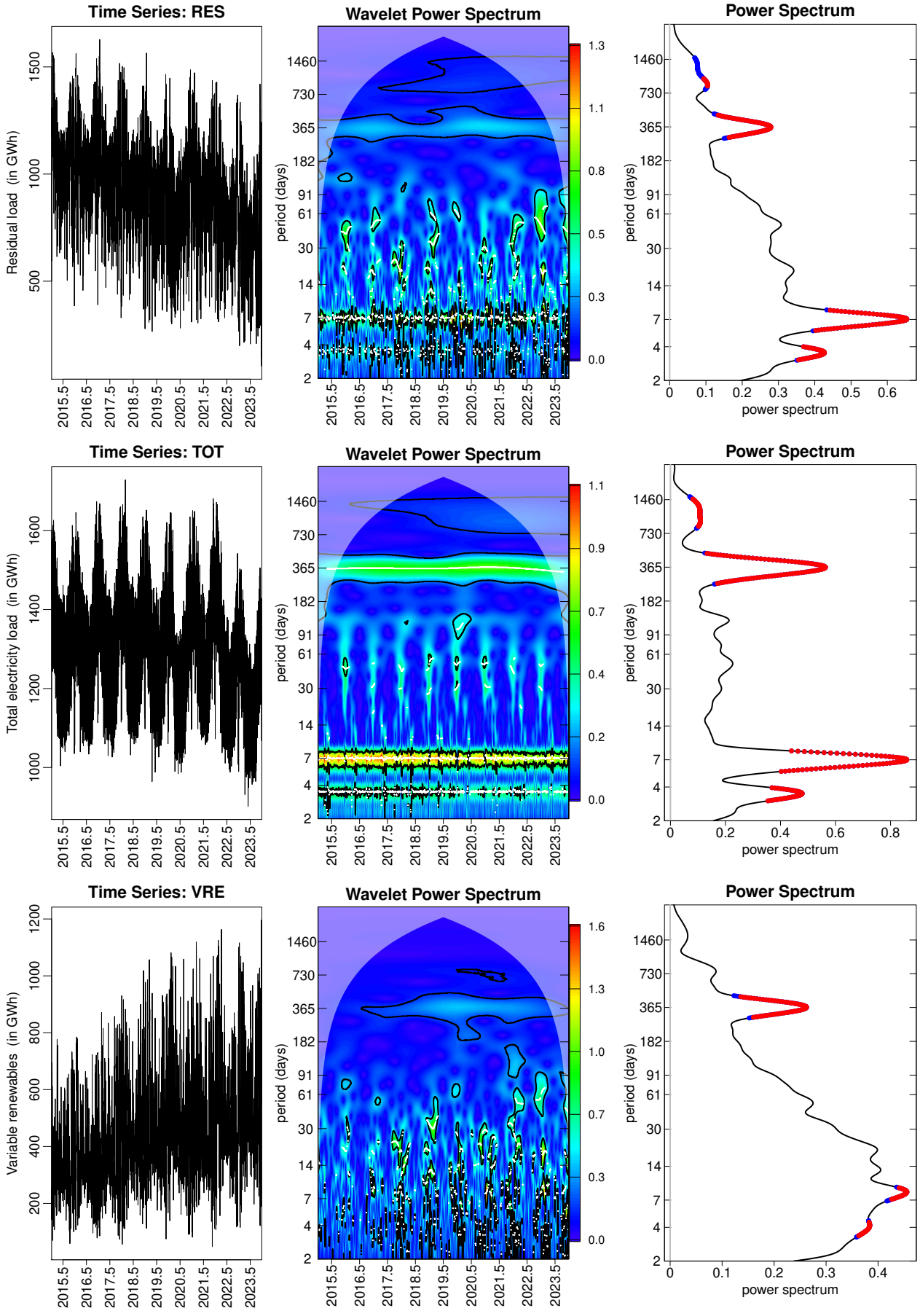


Figure D2: Univariate wavelet results for time series in levels (cont'd)



Affirmation

I hereby declare that the dissertation entitled

"Three Essays in Energy and Environmental Economics"

is my own work. I have only used the sources indicated and have not made unauthorized use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given. I have not presented this thesis or parts thereof to a university as part of an examination or degree.

Place, Date	Signature